LIBRARY ANALYTICS AND BIG DATA

BS(LIS)

Code No. 9219 **Units: 1-9**

Department of Library and Information Sciences
Faculty of Social Sciences and Humanities
ALLAMA IQBAL OPEN UNIVERSITY
ISLAMABAD

LIBRARY ANALYTICS AND BIG DATA

BS(LIS)

Code No. 9219 **Units: 1–9**

AIOU Website: https://aiou.edu.pk

LIS Department Website: https://lis.aiou.edu.pk/

LIS Facebook Page: LIS@AIOU official

DEPARTMENT OF LIBRARY AND INFORMATION SCIENCES
FACULTY OF SOCIAL SCIENCES & HUMANITIES
ALLAMA IQBAL OPEN UNIVERSITY
ISLAMABAD

(All Rights Reserved with the Publisher)

First Edition 2022

Quantity 1000

Price ..... Rs.

Typeset by..... Muhammad Hameed

Printing Incharge..... Dr. Sarmad Iqbal

Printer ..... AIOU-Printing Press, Sector H-8, Islamabad.

Publisher ..... Allama Iqbal Open University, H-8, Islamabad.

COURSE TEAM

Chairman:	Dr. Pervaiz Ahmad Associate Professor
Course Development Coordinator:	Dr. Amjid Khan Assistant Professor
Compiled by:	Dr. Amjid Khan
Reviewed by:	1. Dr. Pervaiz Ahmad 2. Muhammad Jawwad 3. Dr. Muhammad Arif
Edited by:	Humera Ejaz
Layout/Typeset by:	Muhammad Hameed

FOREWORD

Department of Library and Information Sciences was established in 1985 under the flagship of the Faculty of Social Sciences and Humanities intending to produce trained professional manpower. The department is currently offering seven programs from certificate courses to PhD levels for fresh and/or continuing students. The department is supporting the mission of AIOU keeping in view the philosophies of distance and online education. The primary focus of its programs is to provide quality education by targeting the educational needs of the masses at their doorstep across the country.

BS 4-year in Library and Information Sciences (LIS) is a competency-based learning program. The primary aim of this program is to produce knowledgeable and ICT-based skilled professionals. The scheme of study for this program is specially designed on the foundational and advanced courses to provide in-depth knowledge and understanding of the areas of specialization in librarianship. It also focuses on general subjects and theories, principles, and methodologies of related LIS and relevant domains.

This new program has a well-defined level of LIS knowledge and includes courses in general education. The students are expected to advance beyond their higher secondary level and mature and deepen their competencies in communication, mathematics, languages, ICT, general science, and an array of topics of social science through analytical and intellectual scholarship. Moreover, the salient features of this program include practice-based learning to provide students with a platform of practical knowledge of the environment and context, they will face in their professional life.

This program intends to enhance students' abilities in planning and controlling library functions. The program will also produce highly skilled professional human resources to serve libraries, resource access centres, documentation centres, archives, museums, information centres, and LIS schools. Further, it will also help students to improve their knowledge and skills of management, research, technology, advocacy, problem-solving, and decision-making relevant to information work in a rapidly changing environment along with integrity and social responsibility. I welcome you all and wish you good luck with your academic exploration at AIOU!

Prof. Dr. Zia Ul-Qayyum
Vice-Chancellor

PREFACE

This study guide is about the large amounts of data (the seeds of our time) that we are sowing and creating by simple contact with our connected objects or simple use of advanced IT tools and the value generation that we must derive and reap through sophisticated methods and advanced tools.

This book primarily focusses on different types of data and how to handle the "big" amount of structured and unstructured data, which cannot be managed with traditional tools, and deal with its diversity and velocity to generate value. This book is about "big data analytics" which have become one of the most exciting fields of current digital era.

This exciting field opens the way to new opportunities that have significantly changed the business playground. We have noticed that "big" companies such as Google, Facebook, Apple, Amazon, IBM, Netflix and many other companies invest continuously in big data and analytical applications in order to take advantage of every data byte. Many companies have realized that knowledge is power, and to get this power they must gather its source, which is data, and make sense of it.

Thus, the goal of this book is to provide the reader with the different concepts and applications behind big data analytics, those that are necessary and most important to be familiar with the ways in which data analytics process and algorithms work, and how to use them.

This book covers topics on data analytics and metrics; library data analytics, data analytics process, machine learning, supervised versus unsupervised algorithms, data-driven collections management and using qualitative research approaches to transform the library users' experiences.

Every unit of this book is meant for readers who are looking to discover the importance of analytics tools, and who have a critical vision towards how knowledge or this "power" is derived from data. So, if you want to become a data analysis practitioner or a better problem solver, or even if you are considering a career in big data and joining the analytics arena, then this study guide is for you! This study guide is indeed a much-needed publication for the library and education community and a major guide to assist library and information science (LIS) students, LIS professionals, and teaching faculty concerning data analytics and its application in library services and business market.

Additionally, this course has been particularly designed for LIS students with the purpose to prepare them for their future roles in the electronic environment. The expected learning outcomes of this course include a combination of knowledge, values, and skills with a particular emphasis on its use in a professional way. This study guide also provides a much-needed resource for teaching data analytics and big data in the emerging library context to students in higher education.

Prof. Dr Syed Hassan Raza
Dean, Faculty of Social Sciences & Humanities

ACKNOWLEDGEMENTS

First of all, I am extremely grateful to the worthy Vice-Chancellor and the worthy Dean, FSSH for giving me the opportunity to prepare this book. Without their support, this task may not be possible. Further, they have consistently been a source of knowledge, encouragement, benignancy, and much more.

I am highly indebted to my parents, spouse, siblings, and children, who allowed me to utilize family time in completion of this work timely. Their continuous prayers kept me consistent throughout this journey. I would also appreciate the cooperation of my departmental colleagues extended to me whenever required. Special thanks to Academic Planning and Course Production (APCP) and Editing Cell of AIOU for their valued input that paved my path to improve and finish this book in accordance with AIOU standards and guidelines. They have been very kind and supportive as well.

I would also like to thank Print Production Unit (PPU) of AIOU for their support regarding comprehensive formatting of the manuscript and designing an impressive cover and title page. Special thanks also owe to AIOU's library for giving me the relevant resources to complete this task in a befitting manner. I am also thankful to ICT officials for uploading this book on AIOU website. There are many other persons, whose names I could not mention here, but they have been a source of motivation in the whole extent of this pursuit.

Dr Amjid Khan
Assistant Professor, LIS

INTRODUCTION OF THE COURSE

The course has been designed as easy as possible for distance mode of learning and it will help students in completing his/her required course work. The course is of three credit hours and comprises nine units, each unit starts with an introduction which provides an overall overview of that unit. At the start of each unit the objectives of unit show student the basic learning purposes. The rationale behind these objectives is that after reading unit a student should be able to explain, discuss, compare, and analyze the concepts studied in that unit. This study guide specifically structured for students to acquire the skill of self-learning through studying prescribed reading material. Studying all this material is compulsory for successful completion of the course. Recommended readings are listed at the end of each unit. Few self-assessment questions and activities have also been put forth for the students. These questions are meant to facilitate students in understanding and self-assessment that how much they have learned.

For this course, a 6-day workshop at the end of semester will be arranged by the department for learning this course. The participation/attendance in workshop is compulsory (at least 70%). The tutorial classes/meetings are not formal lectures as given in any formal university. These are meant for group and individual discussion with tutor to facilitate students learning. So, before going to attend a tutorial, prepare yourself to discuss course contents with your tutor (attendance in tutorial classes/meetings is non-compulsory).

After completing the study of first 5 units the 'Assignment No. 1' is due. The second assignment that is 'Assignment No. 2' is due after the completion of next 4 units. These two assignments are to be assessed by the relevant tutor/resource person. Students should be very careful while preparing the assignments because these may also be checked with **Turnitin** for plagiarism.

COURSE STUDY PLAN

As you know the course is offered through distance education, so it is organized in a manner to evolve a self-learning process in absence of formal classroom teaching. Although the students can choose their own way of studying the required reading material, but advised to follow the following steps:

Step-1: Thoroughly read description of the course for clear identification of reading material.

Step-2: Carefully read the way the reading material is to be used.

Step-3: Complete the first quick reading of your required study materials.

Step-4: Carefully make the second reading and note down some of the points in notebook, which are not clear and needs fully understanding.

Step-5: Carry out the self-assessment questions with the help of study material and tutor guidance.

Step-6: Revise notes. It is quite possible that many of those points which are not clear and understandable previously become clearer during the process of carrying out self-assessment questions.

Step-7: Make a third and final reading of study material. At this stage, it is advised to keep in view the homework (assignments). These are compulsory for the successful completion of course.

ASSESSMENT/EVALUATION CRITERIA OF STUDENTS' COURSEWORK

Multiple criteria have been adopted to assess students' work for this course, which is as follows:

i. Written examination to be assessed by the AIOU Examination Department, at the end of semester= 70% marks (pass marks 50%). AIOU examination rules will be applied in this regard.
ii. Two assignments and/or equivalent to be assessed by the relevant tutor/resource person= 30% marks (pass marks 50% collectively).

Note: Assignments' submission and getting pass marks is compulsory, the student who will not submit assignments or marked as fail considered FAIL in the course. He/she will get fresh admission in the course; there is no need to sit in the exam.

RECOMMENDED BOOKS

Cooper A. (2012) "What is analytics? Definition and essential characteristics. *CETIS Analytics Series.* 1 (5). 1–10.

Kandasamy, B. P. & Benson, V. (2013). Making the most of big data: Manager's guide to business intelligence success. www.bookboon.com http://93.174.95.29/_ads/EC133CCE54AA14A53992645E9C31BF95

Sedkaoui, S. (2018). Data analytics and big data. John Wiley & Sons.

Showers, B. (Ed.) (2015). Library analytics and metrics: Using data to drive decisions and services. London: Facet Publishing.

OBJECTIVES OF THE COURSE

After reading this course, you will be able to understand, evaluate and describe:

- Data analytics and metrics
- Library data, big data analytics
- Emerging technologies related to big data and data analytics
- Data analytic process
- Text mining
- Data driven collection management
- Issues related to data analytics
- Library Users' experiences

CONTENTS

Unit–1

GETTING THE MEASURE OF ANALYTICS AND METRICS; LIBRARY DATA: BIG AND SMALL

Compiled by: Dr. Amjid Khan

Reviewed by: 1. Dr. Pervaiz Ahmad
2. Muhammad Jawwad
3. Dr. Muhammad Arif

CONTENTS

INTRODUCTION

Data are a collection of facts, such as numbers, words, measurements, observations or even just descriptions of things, which give more information about an individual, an object, or an observation. In this unit discusses data and big data, analytics and metrics and learning analytics with suitable examples. At the end of the unit, self-assessment questions followed by practical activities are given to the students.

OBJECTIVES

After reading this unit, you will be able to understand:

- data and big data
- analytics and metrics
- learning analytics

1.1 INTRODUCTION

Data are a collection of facts, such as numbers, words, measurements, observations or even just descriptions of things, which give more information about an individual, an object, or an observation.

According to the Oxford dictionary, data are defined as:
"The quantities, characters, or symbols on which operations are performed by a computer, which may be stored and transmitted in the form of electrical signals and recorded on magnetic, optical, or mechanical recording media".

In the big data age, we will come across many different types of data, and each of them requires different tools and techniques. Big data is the revolutionary word in today's world because of its influence on several domains. Somehow, the debates that have been emerging over the last few years around big data are very similar to those that took place about the "Web" in the early 1990s. After a long and active discussion phase in the literature, big data entered a phase of use by many organizations and companies.

Businesses and services are adopting analytics to help drive more informed decisions, to gain a better understanding of their customers and users and to make sense of the 'big data' created by all those interactions and actions. Similarly, individuals are increasingly using analytics to help improve their performance and understanding of themselves. The 'quantified self' captures data from activities as diverse as running and sport, through to sleeping and general well-being. These popular applications and services enable the collection and analysis of data to help improve performance in whatever it is you are trying to achieve, whether running, sleeping or productivity at work.

1.2 WHAT IS BIG DATA?

Data exist over time, it is not new, but what makes them so important is the rapid rate and different forms in which they have been produced in recent times, or what brings us to turn: From data to big data. Big data is created digitally and collected automatically. Drawing on an extensive engagement with the literature, big data is:

- huge in *volume*, consisting of terabytes or petabytes of data.
- high in *velocity*, being created in or near real-time.
- diverse in *variety*, being structured and unstructured in nature.
- *exhaustive* in scope, striving to capture entire populations or systems (n = all);
- fine-grained in *resolution* and uniquely *indexical* in identification.

- *relational* in nature, containing common fields that enable the conjoining of different data sets.
- *flexible*, holding the traits of *extensionality* (can add new fields easily) and *scalability* (can expand in size rapidly).

1.3 LIBRARY ANALYTICS

Libraries, along with archives, museums, and galleries, find themselves ideally placed to exploit the full potential of analytics. Libraries, and the cultural sector more generally, have long been familiar with the potential of statistics and data for informing everything from service development to measurement of impact and value (both locally within the institution and nationally – and even internationally). The variety and scope of the data collected and generated by libraries and organizations such as museums and archives is significant: transactional data on catalogue searches, item check-outs, log-ins to online resources and services, swipes through the entrance gates; manually collected statistics on space usage, student satisfaction, external visitors to the library.

The applications of the data are equally varied and overlapping, including management functions (collections development and management, usage statistics), impact (demonstrating value, benchmarking, improving learner outcomes) and improving services and meeting user requirements (recommendation services, collections management/development).

While this diversity in sources and applications is indicative of the importance of data to organizations like libraries, it also highlights the multi-faceted processes and practices for collecting and analyzing the data. These practices are often unique to the local institution and its library and reflect both the accessibility of the data in its local systems and the specific uses and types of data that benefit those institutions and its users. These local variations and challenges would by themselves be sufficient to make this a difficult landscape to traverse, but there are also significant external factors that conspire within the analytics space. Such complications include data access and ownership, formats and standards, privacy, and ethical implications. May be more critically, libraries and other institutions are beginning to question exactly what it is that they are measuring in the first place. There is a need to be clear about what is being measured, and why. Otherwise, there is a very real risk that our measures become too simplistic or, worse, that we are simply measuring the wrong things: ‘we look away from what we are measuring, and why we are measuring, and fixate on the measuring itself.

1.4 LEARNING ANALYTICS

Analytics is the process of developing actionable insight through discovery, modeling and analysis, and interpretation of data. ***Learning Analytics*** is the measurement, collection, analysis, and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs, as defined back in 2011 for the first LAK, this general definition still holds true even as the field has grown. Learning analytics is both an academic field and commercial marketplace which have taken rapid shape over the last decade. As a research and teaching field, *Learning Analytics* sits at the convergence of ***Learning*** (e.g., educational research, learning and assessment sciences, educational technology), ***Analytics*** (e.g., statistics, visualization, computer/data sciences, artificial intelligence), and ***Human-Centered Design*** (e.g. usability, participatory design, sociotechnical systems thinking).

Learning Analytics is concerned with understanding why some students may not be succeeding, what would contribute to their success and how and when interventions might be helpful. The vision is usually to create a more personalized and effective learning experience for students, and even for researchers. The benefits for learners are substantial, and they provide institutions with the opportunity to improve student satisfaction, as well as to enhance completion and retention rates. These are critical success factors for any academic institution. Learning Analytics (LA) is a broad term that spans a broad range of activities: from instructors testing effectiveness of learning approaches, to instructors and advisors determining efficacy of particular learning interventions, to researchers asking basic questions of learning data to gain insights into individual performance or learning strategies, to institutional approaches used for program planning or reporting.

Key Uses. Historically, some of the most common uses of learning analytics is prediction of student academic success, and more specifically, the identification of students who are at risk of failing a course or dropping out of their studies. While it is reasonable that these two problems attracted a lot of attention, learning analytics are far more powerful. The evidence from research and practice shows that there are far more productive and potent ways of using analytics for supporting teaching and learning. Some of the most popular goal of learning analytics includes:

- Supporting student development of lifelong learning skills and strategies.
- Provision of personalized and timely feedback to students regarding their learning.
- Supporting development of important skills such as collaboration, critical thinking, communication, and creativity.
- Develop student awareness by supporting self-reflection.

- Support quality learning and teaching by providing empirical evidence on the success of pedagogical innovations.

The library has a clear role to play in this larger analytics picture, contributing both its data and analytics experiences and its leadership and expertise, in effectively collecting and analyzing data for the benefit of students and in delivering more effective and efficient services.

1.5 SELF-ASSESSMENT QUESTIONS

Q.1 Define data and big data with relevant examples.

Q.2 Describe analytics and metrics with suitable examples.

Q.3 Discuss the role of learning analytics with examples.

Q.4 How learning analytics can be used? Explain.

1.6 ACTIVITIES

- Create library users' data, analyze users' data, and create dashboards that depict meaningful patterns or insights emerging from those analyses.

REFERENCES

Cooper A., (2012). What is analytics? Definition and essential characteristics, CETIS Analytics Series, 1 (5). 1–10.

Kandasamy, B. P. & Benson, V. (2013). Making the most of big data: Manager's guide to business intelligence success. www.bookboon.com http://93.174.95.29/_ads/EC133CCE54AA14A53992645E9C31BF95

Sedkaoui, S. (2018). Data analytics and big data. John Wiley & Sons.

Showers, B. (Ed.) (2015). Library analytics and metrics: Using data to drive decisions and services. London: Facet Publishing.

Unit–2

BUILDING AN UNDERSTANDING OF BIG DATA ANALYTICS

Compiled by: Dr. Amjid Khan

Reviewed by: 1. Dr. Pervaiz Ahmad
2. Muhammad Jawwad
3. Dr. Muhammad Arif

CONTENTS

Page #

INTRODUCTION

Today Big Data draws a lot of attention in the IT world. The rapid rise of the Internet and the digital economy has encouraged an exponential growth in demand for data storage and analytics, and IT department are facing tremendous challenge in protecting and analyzing these increased volumes of information. Hence, this unit highlights the historical perspectives of big data analytics, difference between traditional analytics versus advanced analytics, new statistical and computational paradigms within the big data context and top big and emerging data technologies. At the end of the unit, self-assessment questions followed by practical activities are given to the students.

OBJECTIVES

After reading this unit, you will be able to understand:

- data analytics and historical perspectives of big data analytics
- difference between traditional analytics versus advanced analytics
- new statistical and computational paradigms within the big data context
- top Big Data Technologies
- emerging Big Data Technologies

2.1 WHAT ARE DATA ANALYTICS?

Davenport and Harris define analytics as: "the extensive use of data, statistical and quantitative analysis, explanatory and predictive models, and fact-based management to drive decisions and actions". Analytics team often uses their expertise in statistics, data mining, machine learning and visualization to answer questions and solve problems that management points out. Analytics can also be defined as "a process that involves the use of statistical techniques (measures of central tendency, graphs and so on), information system software (data mining, sorting routines) and operations research methodologies (linear programming) to explore, visualize, discover, and communicate patterns or trends in data".

Business analytics begins with a dataset or commonly with a database. As databases grow, they need to be stored somewhere. Technologies, such as computer and data warehousing, store data. Database storage areas have become so large that a new term was devised to describe them. Business analytics traditionally covers the technologies and application that companies use to collect mostly structured data from their internal legacy systems. These data are then analyzed and mined using statistical methods and well-established techniques classed as data mining and data warehousing.

Generally, there are two main types of analytics:
1. *Descriptive:* which focuses on reporting on what happened in the past;
2. *Predictive:* which uses past data to try and predict future events.

Delen and Demirkan noted that big data adds the ability to perform a third type of analytics, called *perspective analytics*, which combines data from the two previous types and uses real-time external data to recommend an action that must be taken within a certain time to achieve a desired outcome.

Thus, the process of analytics can involve any one of these types; the major components of business analytics include all three used in combination to generate new, unique, and valuable information that can aid business decision-making and organizational performance. In addition, the three types of analytics are applied sequentially (descriptive, then predictive, then prescriptive).

Data processing and analysis, in the present day, are brought together under the notion of "Business Intelligence", due especially to computers' increased processing capabilities. According to Chen *et al*. the term BI became popular in the 1990s, with the term "business analytics" added in the late 2000s to show the importance of analytical capabilities. Analytics has emerged as a catch-all term for a variety of different BI and application-related initiatives. For some, it is the process of analyzing information from a particular domain, such as website analytics. It focuses on knowledge discovery for predictive and descriptive purposes to discover new ideas or to confirm existing ideas. It can be seen from the above definition that data analysis is

a primordial step in the process of knowledge discovery in databases (KDD). This step involves the application of specific algorithms to extract patterns (models) from data. The additional steps are data preparation, data selection, data cleaning, and incorporation of appropriate prior knowledge and proper interpretation of the results of mining. Powerful analytics tools can then be used to process the information gathered in large sets of structured and unstructured data.

2.2 BEFORE AND AFTER BIG DATA ANALYTICS

Before the era of IT tools, company data was mainly in the form of handwritten paper records, which were not easily accessible. More recently, with advanced technology, larger amounts of data can be collected, stored, and reused. And now, a new IT term is coined, *i.e.* the Internet of Things (IoT), in which everything is connected. Therefore, this expands the amount of data, and consequently increases the importance of "Data Analytics".

Data analysis came in the 20th Century when the information age really began. Zhang mentioned in his book *Data Analytics* published in 2017 that the first real data processing machine came during the Second World War. But the advent of the Internet sparked the true revolution in data analysis. Davenport states that company managers have been familiar with using traditional data analysis to support decisions since 1970. Vasarhelyi *et al*. (2015) state that the traditional accounting data in companies were enterprise resource planning (ERP) data, which was acquired manually in transactions. However, the importance of data analysis started in the late 1960s when researchers begin to speak about databases as repositories of data. Codd and his research group at IBM labs applied some mathematical principles and predicate logic to the field of data modeling. Since then, databases and their evolutions have been used as a source of information to query and manipulate data.

In 1974, still at IBM labs, the first language for databases was developed. SEQUEL (Structured English Query Language), later called SQL for copyright issues, was the forerunner of all the query languages, becoming the standard for relational databases in the 1970s and 1980s, and computers could process information, but they were too large and too costly. Only large firms could hope to analyze data with them. Codd was the first to work on data organization by designing database management systems, of relational databases.

Since the 1980s, relational management systems have therefore taken precedence over other systems for the needs of all types of data, first for business and academic systems, then with independent developers for free initiatives or personal use, such as software, websites, etc. Even for small needs, embedded or local systems like SQLite (http://www.sqlite.org/*)* are widely used. The relational model is efficient for a purely transactional use, that which is called "OLTP" (Online Transactional

Processing). A management database, for example, used in ERP, has permanent activity updates and reduced result sets readings. We query the table of filtered invoices for a customer, which returns a dozen lines; we request the table of payments to verify that this customer is solvent; if so, we add an invoice with lines of invoices, and for each product added, it decrements its stock in the product table. All these operations have limited scope in tables whose cardinality (the number of lines) can be otherwise important. But, because of a good data modeling, each of these operations is optimized. In an ERP, there is a complete analysis of sales trends by product category, branch, department, month and by customer types, calculating developments to determine which product categories are changing, in which region and for which customer, etc.

In this kind of query, or what we call online analytical processing ("OLAP"), which has to cover a large part of the data to calculate aggregates, the relational model and the query optimizers of the databases cannot respond satisfactorily to the need.

The OLAP model was created due to increased aggregated and historical data storage and global query requirements on these large volumes for analytical purposes. This is called BI. This model, which has also been formalized by Codd, prefigures the big data phenomenon. In recent years, with the advent of Web 2.0 and the semantic web era, data analysis has become very important, replacing the traditional storing systems in many applications. They now represent the new technology for knowledge representation, data storage and information sharing.

2.3 DIFFERENCE BETWEEN TRADITIONAL ANALYTICS VERSUS ADVANCED ANALYTICS

- Traditional analytics (descriptive) provides a general summary of data while advanced analytics deliver deeper data knowledge.
- Traditional analytics mines past data to report, visualize and understand what has already happened. While modern analytics leverages past data to understand why something happened? Or to predict what will happen in the future across various scenarios.
- Advanced analytics determines which decision and/or action will produce the most effective result against a specific set of objectives and constraints.
- New analytics approaches in the big data age combine predictive and prescriptive analytics to predict what will happen and how to make it happen. Analytics' uses and applications improve the efficiency of decision-making processes and generate value.
- Advanced analytics can range from historical reporting to real-time decision support for organizations based on future predictions.

2.4 NEW STATISTICAL AND COMPUTATIONAL PARADIGM WITHIN THE BIG DATA CONTEXT

Statistics is the traditional field that deals with the quantification, collection, analysis, interpretation, and evaluation of data. The development of new statistical methods is an interdisciplinary field that draws on computer sciences, artificial intelligence, machine learning, and visualization models and so on. There are several methods that have recently been developed and are feasible for statistical inference of big data and workable on parallel machines, including the bag of little bootstraps, aggregated estimation equation and so on. Each method was developed to find and design tools that explicitly reveal tradeoffs relating complexity, risk and time.

Concerning statistical methods, the literature summarizes the change in two points:

- *the new approaches are on the crossroads of IT tools and statistics:* this concerns machine learning, where algorithms generate, more or less alone models on large amounts of data;
- *these methods are not new because machine learning dated from the 1960s:* this return to the center stage is since these techniques work especially well on high amounts of information.

Big data poses new challenges to statisticians both in terms of theory and application. Some of the challenges include size, scalability of statistical computation methods, non-random data, assessing uncertainty, sampling, modeling relationships, mixture data, real-time analysis of streaming data, statistical analysis with multiple kinds of data, data quality and complexity, protecting, privacy and confidentiality, high dimensional data, etc.

As the volume of data grows, so do the requirements for more advanced data warehouses and dispersed cloud-based databases. Good theoretical insights and models of the subject discipline would be useful in identifying the "payoff relevance" of data for predictive purposes.

The notion of making sense of big data has been expressed in many ways, including data mining, knowledge extraction, information discovery, information harvesting, data archaeology and data pattern processing.

2.5 TOP 5 BIG DATA TECHNOLOGIES

2.5.1 Hadoop Ecosystem

Hadoop Framework was developed to store and process data with a simple programming model in a distributed data processing environment. The data present on different high-speed and low-expense machines can be stored and analyzed. Enterprises have widely adopted Hadoop as Big Data Technologies for their data

warehouse needs in the past year. The trend seems to continue and grow in the coming year as well. Companies that have not explored Hadoop so far will most likely see its advantages and applications.

2.5.2 Artificial Intelligence

Artificial Intelligence is a broad bandwidth of computer technology that deals with the development of intelligent machines capable of carrying out different tasks typically requiring human intelligence. AI is developing fast from Apple's Siri to self-driving cars. As an interdisciplinary branch of science, it considers several approaches such as increased Machine Learning and Deep Learning to make a remarkable shift in most tech industries. AI is revolutionizing the existing Big Data Technologies.

2.5.3 NoSQL Database

NoSQL includes a wide variety of different Big Data Technologies in the database, which are developed to design modern applications. It shows a non-SQL or non-relational database providing a method for data acquisition and recovery. They are used in Web and Big Data Analytics in real-time. It stores unstructured data and offers faster performance and flexibility while addressing various data types—for example, MongoDB, Redis and Cassandra. It provides design integrity, easier horizontal scaling, and control over opportunities in a range of devices. It uses data structures that are different from those concerning databases by default, which speeds up NoSQL calculations. Facebook, Google, Twitter, and similar companies store user data terabytes daily.

2.5.4 R Programming

R is one of the open-source Big Data Technologies and programming languages. The free software is widely used for statistical computing, visualization, unified development environments such as Eclipse and Visual Studio assistance communication. According to experts, it has been the world's leading language. The system is also widely used by data miners and statisticians to develop statistical software and mainly data analysis.

2.5.5 Data Lakes

Data Lakes means a consolidated repository for storage of all data formats at all levels in terms of structural and unstructured data. Data can be saved during Data accumulation as is without being transformed into structured data. It enables performing numerous types of Data analysis from dashboards and Data visualization to Big Data transformation in real-time for better business interference. Businesses that use Data Lakes stay ahead in the game from their competitors and carry out new analytics, such as Machine Learning, through new log file sources, data from social media and click-streaming. This Big Data technology helps enterprises respond to better business growth opportunities by understanding and engaging clients, sustaining productivity, active device maintenance, and familiar decision-making to better business growth opportunities.

2.6 EMERGING BIG DATA TECHNOLOGIES

2.6.1 Tensor Flow

Tensor Flow has a robust, scalable ecosystem of resources, tools, and libraries for researchers, allowing them to create and deploy powerful Machine Learning applications quickly.

2.6.2 Beam

Apache Beam offers a compact API layout to create sophisticated Parallel Data Processing pipelines through various Execution Engines or Runners. Apache Software Foundation developed these tools for Big Data in the year 2016.

2.6.3 Docker

Docker is one of the tools for Big Data that makes the development, deployment and running of container applications simpler. Containers help developers stack an application with all the components they need, such as libraries and other dependencies.

2.6.4 Airflow

Apache Airflow is a Process Management and Scheduling System for the management of data pipelines. Airflow utilizes job workflows made up of DAGs (Directed Acyclic Graphs) tasks. The code description of workflows makes it easy to manage, validate and version a large amount of Data.

2.6.5 Kubernetes

Kubernetes is one of the open-source tools for Big Data developed by Google for VENDOR-agnostic cluster and container management. It offers a platform for the automation, deployment, escalation, and execution of container systems through host clusters.

2.6.6 Blockchain

Blockchain is the Big Data technology that carries a unique data safe feature in the digital Bitcoin currency so that it is not deleted or modified after the fact is written. It’s a highly secured environment and an outstanding option for numerous Big Data applications in various industries like baking, finance, insurance, medical and retail, to name a few.

2.7 SELF-ASSESSMENT QUESTIONS

Q.1 Describe data analytic. Also write a note on the historical perspectives of data analytics with relevant examples.

Q.2 What is the difference between traditional analytics versus advanced analytics? Discuss with examples.

Q.3 What are the new statistical and computational paradigms within the big data context? Explain.

Q.4 Write a comprehensive note on top and emerging big data technologies.

2.8 ACTIVITIES

- Enlist new emerging big data analytics tools and highlights its main features.

REFERENCES

Cooper A., (2012). What is analytics? Definition and essential characteristics, CETIS Analytics Series, 1 (5). 1–10.

Kandasamy, B. P. & Benson, V. (2013). Making the most of big data: Manager's guide to business intelligence success. www.bookboon.com.http://93.174.95.29/_ads/EC133CCE54AA14A53992645E9C31BF95.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Hung Byers, A. (2011). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute.

Sedkaoui, S. (2018). Data analytics and big data. John Wiley & Sons.

Vasarhelyi, M. A., Kogan, A., & Tuttle, B. M. (2015). Big data in accounting: An overview. *Accounting Horizons*, *29*(2), 381–396.

Unit–3

WHY DATA ANALYTICS AND WHEN CAN WE USE IT?

Compiled by: Dr. Amjid Khan

Reviewed by: 1. Dr. Pervaiz Ahmad
2. Muhammad Jawwad
3. Dr. Muhammad Arif

CONTENTS

INTRODUCTION

The process of gathering, processing, and interpreting information is not limited to defining ideas, but also consists of materializing them to ensure improved knowledge production that leads to value creation. This unit focuses on the use of data analytics and real time data analysis, issues in data analytics and analytics culture within organizations/companies. At the end of the unit, self-assessment questions followed by practical activities are given to the students.

OBJECTIVES

After reading this unit, you will be able to explain:

- the use of data analytics and real time data analysis
- challenges/issues in data analytics
- analytics culture within organizations/companies
- examples of using data analytics by organizations

3.1 INTRODUCTION

The reason organizations are collecting and storing more data than ever before is because their business depends on it. The type of information being created is no more traditional database-driven data referred to as structured data rather it is data that include documents, images, audio, video, and social media contents known as unstructured data or Big Data. Big Data Analytics is a way of extracting value from these huge volumes of information, and it drives new market opportunities and maximizes customer retention. The process of gathering, processing, and interpreting information is not limited to defining ideas, but also consists of materializing them to ensure improved knowledge production that leads to value creation. Big data analytics can be defined as a new discipline born from the mixture of statistics, computer sciences and business. It allows each company to optimize its operation and strategy. The data must be analyzed and reviewed to be able to add value, especially when the aim is to optimize methods and operations.

The increase in data produced by companies, individuals, scientists, and public officials, coupled with the development of IT tools, offers new analytical perspectives. Analysis of big data requires an investment in computing architecture to store, manage, analyze, and visualize an enormous amount of data. If Facebook, Google, Twitter, LinkedIn, Amazon, Apple, Netflix, Nike, and many other data-driven business models exist, it is because of the advantages generated by big data and analytics. Every second, visitors interact with interconnected objects and leave behind a tremendous amount of data that companies can then use to create tailor-made experiences. Faced with such a challenge, both make sure that the technologies used can correctly handle this volume of data. Big data and the use of data analytics are being adopted more frequently, especially in companies that are looking for new methods to develop smarter capabilities and tackle challenges in dynamic processes. Big data and profit are closely intertwined. The correlation identified by Google between a handful of search terms and the flu is the result of testing 450 million mathematical models. Big data analytics has become an essential requisite to run most businesses. Integrating big data analytics can generate many advantages for the companies, such as:

***Supporting decision*:** companies can make use of the vast amount of data relevant to their business. Therefore, they would need to filter the data according to their specific needs and derive meaning from the data that fits them best. This will not only widen their understanding of their own domain but will also facilitate better decision-making, which in turn will improve operational efficiencies.

***Cost reduction*:** it has been found that big data can be extremely instrumental in augmenting the existing architectures of companies.

Additionally, when more accurate decisions are taken, the possibility of incurring losses also gets alleviated. Therefore, with the correct use of analytics, businesses can be successful in cutting down their operational costs, which is typically one of the biggest challenges.

***Customer insights*:** the growth of any company depends on how to consider the preferences, likes, tastes, etc. of their customers in the design of their products and services. Big data analytics can help companies to gain access to the required and relevant information. For example, social media presents a great tool to acquire and assimilate enormous volumes of customer insights and can be used effectively to collect data for this purpose.

Open data *uses* over the last year, there has been an increase in the perceived use of open data to build new products and services. Open data, in addition to its economic potential and the creation of new activities it entails, also falls within a domain of philosophy or ethics. It belongs to individuals and can be used to encrypt their behavior. The culture of this phenomenon builds on the availability of data for a communication orientation.

Combined with advanced analysis methods, new explanations can be provided for several phenomena. The analyzed data allow companies to obtain strategic advantages by considering a greater number of data, by improving existing ones (optimization and cost reduction, productivity gains, etc.), by creating new ones, more targeted products, or to improve the customer experience. Indeed, better knowledge of new needs is clearly an asset. Because the data state a lot about customer preferences, they represent business and marketing issues for the company.

Data mining, for example, is a technique that can extract and analyze big data to extract relevant information. Simply put, data mining software processes the huge amount of data to highlight trends, models, and correlations. For example, if your company sells air conditioners, you may have noticed that sales increase during the summer. Data mining makes it possible to be much more precise and to highlight that the sales of these kinds of products increase a few days after a heat wave, since the average maximum temperature of these days exceeds 28 degrees.

This makes it possible to accurately predict when the demand will increase (or not), to accordingly adapt the rate of production and the supply chain or to launch an advertising campaign at the right time. The models and correlations established by data mining not only make it possible to understand the present, but also to anticipate behaviors.

In this, it serves as a basis for machine learning. Industrial maintenance, the customization of offers, energy efficiency, preventive medicine and the autonomous car are all examples of the different applications of big data analytics.

The profound transformations engendered by large-scale data processing capacity can redefine the boundaries between different sectors and companies that will seize the opportunity. Cities and even entire countries also benefit from data analysis. Thus, real-time resource management is now possible.

3.2 WHEN REAL TIME MAKES THE DIFFERENCE

The challenge is not only to collect the data, but also to exploit process and analyze them better, in such a way that allows businesses to generate knowledge to upgrade the process of decision making and achieve higher performance. Beyond the advent of ICT and increased data production, dissemination and processing speeds, another element has recently become critically important: "time". The importance of time carries with it a notion of information circulation speed. So, big data adds an unprecedented dimension: exploiting the profusion of huge volumes of data with the finest level of detail and often the shortest lifetime (instantaneity). And it will be more performant if it becomes possible to analyze the data in real time. Traditional data processing architectures know how to collect data before transforming and then analyzing them. These three operations are globally performed one after another. Today several elements are combined to define an integrated system of "big data injection in real time", which is about to transform business operations.

Real-time analytics was growing at a rapid pace and is now poised to reach an inflection point. This evolution could not come at a more opportune moment. In real-time analytics, one analyzes and visualizes data in real time. For example, now, with the power of real-time data processing, a health service provider can continually monitor patients at risk. By combining the real-time data recorded by several connected object to monitor medical symptoms with information in medical records, analysis tools can alert health professionals if proactive action is urgent for the patient.

3.3 WHAT SHOULD DATA ANALYTICS ADDRESS?

The analysis of a larger amount of data in real time is likely to improve and accelerate decisions in multiple sectors, from finance to health, both including research. The considerable increase in the volume and diversity of digital data generated, coupled with big data technologies, offers significant opportunities for value creation.

This value cannot be reduced to simply what we can solve or improve, but rather it knows what the new potential discoveries are that may arise from cross-exchanges and correlations. This leads us to say that new data processing tools are now necessary, as are methods capable of combining thousands of datasets. It is the use of data that empowers decision-making. Being increasingly aware of the

importance of data and information, companies are pressed to rethink the way to "manage", to enrich and to benefit from them. This causes two main challenges as follows: Big data contains invisible models, which must be viewed using tools and analytical techniques. The knowledge gained should be used at the right time in the right context and with the right approach.

Capturing, managing, combining, securing, and always taking advantage of a huge amount of data are much more complicated than the simple data storage problem.

As large datasets are currently available from a wealth of different sources, companies are looking to use these resources to promote innovation, customer loyalty and increase operational efficiency. At the same time, they are contested for their end use, which requires a greater capacity to collect, analyze and manage the growing amount of data but also ensure its security. This highlights that it is not merely the existence of large amounts of data that is creating new security challenges. Data exploration and analysis turned into a difficult problem in many sectors in the case of big data.

Let us think about big data in network cybersecurity, an important problem. Governments, corporations, financial institutions, hospitals and other business collect process and store confidential information on computers and transmit that data across networks or other computers. With large and complex data, computation became difficult to be handled by the traditional data processing applications and triggered the development of big data applications. If big data are combined with predictive analytics, they produce a challenge for many industries. The combination results in the exploration of these four areas are as follows:

- Calculate the risks on large portfolios.
- Detect, prevent and re-audit financial fraud.
- Improve delinquent collections.
- Execute high-value marketing campaigns.

The main challenges associated with the development and deployment of big data analytics are as follows:

***The heterogeneity of data streams*:** dealing with semantic interoperability of diverse data streams requires techniques beyond the homogenization of data formats. Big data streams tend to be multimodal and heterogeneous in terms of their formats, semantics and velocities. Hence, data analytics typically expose variety and veracity. Big data technologies provide the means for dealing with this heterogeneity in the scope of operationalized applications.

Data quality: the nature of data available can be classified as noisy and incomplete, which creates uncertainty in the scope of the data analytics process. Statistical and probabilistic approaches must therefore be employed to consider the noisy nature

of data. Also, data can be typically associated with varying reliability, which should be considered in the scope of their integration in the analytical approach.

The real-time nature of big datasets: big data feature high velocities and for several applications must be processed nearly in real time. Hence, data analytics can greatly benefit from data streaming platforms, which are part of the big data ecosystem. IT, the Internet and several connected objects typically provide high-velocity data, which can be in several cases controlled by focusing only on changes in data patterns and reports, rather than dealing with all the observations that stem from connected objects.

The time and location dependencies *of big data*: IoT data come with temporal and spatial information, which is directly associated with their business value in an analytics application context. Hence, data analytics methods must in several cases process data in a timely fashion and from proper locations. Cloud computing techniques (including edge computing architectures) can greatly facilitate timely processing of data from several locations in the case of large-scale deployments. Note also that the temporal dimensions of big data can serve as a basis for dynamically selecting and filtering streams toward analytics tools for certain timelines and locations.

Privacy and security sensitivity**:** big data are typically associated with stringent security requirements and privacy sensitivities, especially in the case of IoT applications that involve the collection and processing of personal data. Hence, advanced analytics need to be supported by privacy preservation techniques, such as the anonymization of personal data, as well as techniques for encrypted and secure data storage.

Data bias**:** as in most data mining problems, big datasets can lead to biased processing and hence a thorough understanding and scrutiny of both training and test datasets is required prior to their operationalized deployment. Note that the specification and deployment of IoT analytics systems entails techniques like those deployed in classical data mining problems, including the understanding and the preparation of data, the testing of the analytics techniques and ultimately the development and deployment of a system that yields the desired performance and efficiency.

Data acquisition and recording**:** it is critical to capture the context in which data have been generated, to be able to filter out non relevant data and to compress data, to automatically generate metadata supporting rich data description and to track and record provenance.

Information extraction and cleaning**:** data may have to be transformed to extract information from it and express this information in a form that is suitable for

analysis. Data may also be of poor quality and/or uncertain. Data cleaning and data quality verification is thus critical.

***Data integration, aggregation, and representation*:** data can be very heterogeneous and may have different metadata. Data integration, even in more conventional cases, requires huge human efforts. Novel approaches that can improve the automation of data integration is critical as manual approaches will not scale to what is required for big data. Also, different data aggregation and representation strategies may be needed for different data analysis tasks.

***Query processing and analysis*:** methods suitable for big data need to be able to deal with noisy, dynamic, heterogeneous, untrustworthy data and data characterized by complex relations. However, despite these difficulties, big data, even if noisy and uncertain, can be more valuable for identifying more reliable hidden patterns and knowledge compared to tiny samples of good data. Also the (often redundant) relationships existing among data can represent an opportunity to cross-check data and thus improve data trustworthiness. Supporting query processing and data analysis requires scalable mining algorithms and powerful computing infrastructures.

***Interpretation and visualization*:** analysis results extracted from big data needs to be interpreted by decision-makers and this may require the users to be able to analyze the assumptions at each stage of data processing and possibly retrace the analysis. Rich provenance is critical in this respect. This process is supported by cloud computing and computational tools, including data mining, statistical computing, and scalable databases technology.

3.4 ANALYTICS CULTURE WITHIN ORGANIZATIONS/ COMPANIES

In all sectors, the uses of big data analytics testify to the effectiveness of this technology. Hospitals and healthcare practitioners are collecting information about patients as they can predict epidemics and design new treatments that can reduce waste and improve service delivery. Another potential benefit of big data is that it can provide more regular and timely information on interesting patterns, such as early indicators of epidemics, economic upturns or downturns, e.g. Google's flu indicators despite its problems, unemployment or housing boom etc., because of the lower unit cost of acquiring big data sources than the traditional direct data collection methods used by NSOs.

In the dynamic environment, records and management systems are independently maintained by education institutions, libraries, and books whilst data are not readily accessible in a centralized position. Big data is being created due to digitalization of libraries and this has imposed limitations to researchers, educationists, scholars, and

policy maker's efforts in improving the quality and efficiency. As a result, serving the users with books and articles that are in line with their interests is a great challenge.

Prediction approaches and analytics methods have strong roles in the creation of social and economic opportunity. But they are not only a brand that bears the large companies. Analytical practices also feature in every entrepreneur's toolkit.

Many successful entrepreneurs' experiences support analytics as a core capability of their startups. These include Sergey Brin and Larry Page of Google, Jeff Bezos of Amazon.com, Michael Bloomberg of Bloomberg LP and Reed Hastings of Netflix. They have seen the potential of using analytics not only to differentiate but also to innovate their business models.

The analysis of big data is not only a matter of solving computational problems, even if those working on big data come from the natural sciences or computational fields. Rather, expertly analyzing big data also requires thoughtful measurement, careful research design and the creative deployment of statistical techniques.

Indeed, massive datasets will require the full range of statistical methodology to be brought to bear for assertions of knowledge based on massive data analysis to be reliable. Following a period when the main issue is how to organize and structure databases? The question now is what to do? What kind of analysis must be applied, adopted, and developed to value and support decision-making? In another words, how should analytical approaches be designed to be scalable computationally to the massive datasets? Then, to capitalize on its potential, companies must put data analytics at the center of their strategy. What are truly necessary are excellent analytic and soft skills, a capacity to understand and manipulate large sets of data, and the capacity to interpret and apply the results. But they also need to establish clear guidelines for data integrity and security, as digital ecosystems can only function efficiently if all parties involved can trust in the security of their data and communication. So, there are many varieties of factors that have contributed to the growing use of big data analytics and its applications. Generally, these factors have also spurred adoption of "Artificial Intelligence" (AI) and "Machine learning" (ML) in the business context. AI and ML are being adopted widely in sectors such as oil, industry, health, manufacturing, marketing, and many other areas. These include availability of computing power owing to faster processor speeds, lower hardware costs and better access to computing power via cloud services. In the age of the data revolution, analytics is empowering companies to take a pragmatic, evidence-based approach to all aspects of their business, including communications and marketing, operations, transportation and logistics, cyber security, and risk management.

3.5 EXAMPLES OF DATA ANALYTICS APPLICATIONS

The new analytical power is seen as an opportunity to invent and explore new methods, which can detect correlations between the quantities of available data. Cukier and Mayer-Schoenberger see a paradigmatic change in the statistical handling of large data "using great volumes of information require three profound changes in how we approach data".

The first is to collect and use a lot of data rather than settle for small amounts or samples as statisticians have done for well over a century. ***The second*** is to shed our preference for highly curated and pristine data and accept messiness: in an increasing number of situations, a bit of inaccuracy can be tolerated, because the benefits of using vastly more data of variable quality outweigh the costs of using smaller amounts of very exact data. ***Third,*** in many instances, we will need to give up our quest to discover the cause of things, in return for accepting correlations. With big data, instead of trying to understand precisely why an engine breaks down or why a drug's side effect disappears, researchers can instead collect and analyze massive quantities of information about such events and everything that is associated with them, looking for patterns that might help predict future occurrences. Big data helps answer what, not why, and often that's good enough".

There are *five* broad ways in which using big data can create value.
First, big data can unlock significant value by making information transparent and usable at much higher frequency.

Second, as organizations create and store more transactional data in digital form, they can collect more accurate and detailed performance information on everything from product inventories to sick days, and therefore exposes variability and boost performance.

Third, big data allows ever narrower segmentation of customers and therefore much more precisely tailored products or services.

Fourth, sophisticated analytics can substantially improve decision-making.

Finally, big data can be used to improve the development of the next generation of products and services".

Big data analysis is essential when organizations want to engage in predictive analysis, natural language processing, image analysis or advanced statistical techniques such as discrete choice modeling and mathematical optimization, or even if they want to mash up unstructured content and analyze it with their BI. Companies will be able to suggest data management for decision making.

The new analytical power is seen as an opportunity to invent and explore new methods, which can detect correlations between the quantities of available data. Large companies are increasingly using big data analytics to improve their business. Also, Shell uses big data and industrial IoT to develop a "data driven oil field" that brings down the cost of production, monitors equipment in real time, manages cyber risks and increases efficiency of transport, refinement, and distribution. Many other companies use data analytics.

3.6 SELF-ASSESSMENT QUESTIONS

Q. 1 Why data analytics should be used in today's complex information environment? Justify your answer with examples.

Q. 2 Describe real time data analytics with relevant examples.

Q. 3 Write a comprehensive note on the challenges during data analytics.

Q. 4 How organizations adopt data analytics applications? Explain with examples.

3.7 ACTIVITIES

Suppose 500 different users daily visit an academic library website to search/access required information. How would you use data analytics to explore the main purpose of their visit to library?

REFERENCES

Cooper A., (2012). What is analytics? Definition and essential characteristics, CETIS Analytics Series, 1 (5). 1–10.

Kandasamy, B. P. & Benson, V. (2013). Making the most of big data: Manager's guide to business intelligence success. www.bookboon.com. http://93.174.95.29/_ads/EC133CCE54AA14A53992645E9C31BF95

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Hung Byers, A. (2011). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute.

Sedkaoui, S. (2018). Data analytics and big data. John Wiley & Sons.

Vasarhelyi, M. A., Kogan, A., & Tuttle, B. M. (2015). Big data in accounting: An overview. *Accounting Horizons*, *29*(2), 381–396.

Unit–4

DATA ANALYTICS PROCESS

Compiled by: Dr. Amjid Khan

Reviewed by: 1. Dr. Pervaiz Ahmad
2. Muhammad Jawwad
3. Dr. Muhammad Arif

CONTENTS

INTRODUCTION

Data analytics is a process of analyzing raw datasets to derive a conclusion regarding the information they hold. Data analytics processes and techniques may use applications operating on machine learning algorithms, simulation, and automated systems. This unit primarily focuses on data analytics process, big data analytics tools, data mining and text mining techniques. It also covers topics on Machine Learning (ML), database management systems, data streams management systems. At the end of the unit, self-assessment questions followed by practical activities are given to the students.

OBJECTIVES

After reading this unit, you will be able to explain:

- data analytics process and big data analytics tools
- data mining and text mining
- machine learning (ML)
- database management systems
- data streams management systems
- building a model

4.1 INTRODUCTION

Analyzed data are no longer necessarily structured in the same way as in traditional analysis, but can now be text, images, multimedia content, digital traces, connected objects, etc. Before tackling the subject of a data analytics process, some points deserve to be relieved.

Everything in big data analytics begins with a clear problem statement. The success of an analytics approach cannot be possible without the clarification of what you want to achieve. This is not just valid in the context of big data, but in all areas. We must clearly define what we want before undertaking anything. This means knowing what we are trying to achieve, what is needed, and why and what level of accuracy is acceptable and actionable. What should be done in this phase is to explore all possible paths to recover the data to identify all the variables that affect, directly or indirectly, the phenomenon that interests us. In other words, how do we make new opportunities from these data? Which data should we select for the analysis? And how do we efficiently apply analytical techniques to generate value? What new insights can I expect? Where does the greatest global potential of big data lie? How will these insights help me? Having the ability to think critically allows you to understand that big data opportunities are not in the volume of data, but in the digital transformation of your business processes. Understanding the basics: identify what we already know and what we have yet to find out

The data are mostly available, but often scattered in several computer tools. An important procedure is to understand the data that will be collected and then analyzed. The idea is that the better your understanding of your data, the better you will be able to use them wisely during the modeling phase. This aims to precisely determine where we should look for the data and which data we should analyze and identify the quality of the data available as well as link the data and their meaning from a business perspective. That means understanding first what we can do with these data before exploring them. This includes some basic knowledge about the methods that will be used and complexities involved.

It seems obvious because data are the main raw material of an efficient data analysis process. So, if you do not understand the nature of the data related to the problem you are trying to solve, consider that you will not be able to solve it. Formulating some business questions to develop a method is important, such as: which sources do they use? What data to collect?

Why these data? To do what? What answers to expect? How much data should be processed? Should we do analysis in real time or periodically? Therefore, to

understand the context of the target problem, you must play, in some ways, a role of a detective. This can allow you to discover and understand the different elements related to it and determine the tools you need. It is therefore essential to have at least a basic notion of statistics and mathematics to determine the right analysis technique according to the nature of each data. That also means a significant part of identifying the technologies that will be most relevant for managing the volume and flow of data.

4.2 DEFINING THE TASKS TO BE ACCOMPLISHED

The specific task to be accomplished corresponds to the problem we are trying to solve by modeling the situation. We can distinguish several cases that often recur in a business environment, such as product recommendations. Each case will translate differently and will of course require the choice of different techniques and algorithms.

4.3 WHICH TECHNOLOGY TO ADOPT?

For data analytics, preference is given mainly to computer languages, which are standardized for data analysis and information extraction. To meet these information-sharing developments, we need tools across the board to help. We need infrastructure and technologies that accommodate ultrafast data capture and processing.

Devices, networks, central processing, and software are used to help us discover and harness new opportunities. The key elements to identify are as follows, considering the growing advanced analytics tools: which technology to adopt? Why should enterprises adopt this technology? Reporting tools, dashboards, data visualization tools, self-service, data warehouses and real-time data analytics are the most-used technologies in business intelligence (BI).

The processing stages that apply to BI applications also apply to big data but demand extra technological effort to enable the complete process of data capturing, storage, search, sharing, analytics and visualization to occur smoothly. Understanding data analytics is good but knowing how to use it is better! (What skills do you need)?

When working with big data, do not only focus on the technological issues, but also on how we can turn that data into patterns. An algorithm is a “black box”; a user can introduce data (inputs) and they will obtain the results (outputs). How the algorithm works is not the user’s business? It is like when someone drives a car

without having any idea about its mechanisms, because knowledge is different from know-how.

In addition, knowing some analytics methods (decision tree, K-means, etc.) can be a real asset. Since these different techniques can be directly implemented using software (SAS, R, Python, etc.), it is not necessary to know how their algorithms work. The important thing is to understand how they work in general terms and to know which method is most relevant depending on the situation. Selecting the right method for the data that you have is a very important step.

4.4 WHAT DOES THE DATA PROJECT COST AND HOW WILL IT PAY OFF IN TIME?

First, you need to determine whether the targeted data give you a return on the investment made by collecting and storing them. Investing in data whose processing cost would be higher than their probable value is indeed to be avoided. Then, you must also have an idea about some issues such as: how much data is needed for each step? What is it meant to achieve? Any presentation of data analysis results must be accompanied by a summary of the resources committed (investments, etc.) and earnings expectations directly related to the analysis produced by these resources.

What will it mean to you once you find out?
Knowing something about everything equips you to understand the context and can extract the value from data. The key is to think broadly about how to transform data into a form which would help us to find valuable tendencies and interrelationships. Seek the correct data to answer a given question. To think like a data scientist or to be a data scientist means thinking about the "meaningfulness" of data, and thus its practice.

Next steps: do you have an idea about a "secret sauce"? Big data analysis is a complex process, and we need to focus on extracting useful knowledge from many different types of data (structured, semi structured, and unstructured). This process involves multiple distinct phases that include data acquisition and recording; information extraction and cleaning; data integration, aggregation, and representation; query processing; data modeling and analysis and interpretation.

First phase: find the data (data collection). Data are collected and enriched with the support of advanced technology (sensors, etc.). Moreover, the data are validated in terms of their format and source of origin. Also, they are validated in terms of their integrity, accuracy, and consistency.

Second phase: construct the data (data preparation): The data preparation phase groups the activities related to the construction of a dataset to be analyzed that is made from the raw data. This includes the classification of data according to selected criteria, the cleaning of data, and especially their recoding to make them compatible with the algorithms that will be used. It must also be ensured that the data are consistent, without missing values for example. Then, all these data must be centralized in a database. Rest assured that you do not need to know the most complex algorithms but you must have a good knowledge about the data and prepare the ground with upstream processing. The important thing is to prepare the ground for the next steps, which will be greatly simplified if this tedious work is done well upstream.

Third phase: go to exploration and modeling (data analysis): The collected, cleaned, and prepared data can now be explored. Finally, you can enter the most interesting phase of the analytics process, i.e., the creation of the analytical model associated with the data we are interested in. This step allows us to better understand the different behaviors and to understand the underlying phenomenon. Feel free to display all kinds of graphs, compare different variables to each other, test correlation hypotheses, clustering, etc. At the end, you will be able to:

Propose several hypotheses about the causes underlying the generation of the dataset: for example, you will be able to understand if there is really a relationship between two variables X and Y. Build several possible statistical modeling paths that can help in solving the problem statement. Introduce, if necessary, new sources of data that would help you to better understand the problem.

Fourth phase: evaluate and interpret the results (evaluation and interpretation) Before operationalizing the model, you need to evaluate the quality of the model, i.e., its ability to accurately represent our case study, or at least its ability to solve our problem statement. Good results require an effective strategy of data collection, preparation, and analysis. The evaluation verifies the model obtained to ensure that it meets the objectives formulated at the beginning of the process. It also contributes to the decision of the deployment of the model or, if necessary, to its improvement. At this stage, the robustness and accuracy of the models obtained are tested.

Fifth phase*:* transform data into actionable knowledge (deploy the model)
As part of this phase, the analytics techniques, and models identified in the previous phase are becoming operational. This phase also ensures the visualization of the data/knowledge according to the needs of the situation. This is the final phase of the process.

It consists of a production run for the obtained models' end users. Its aims to mold the knowledge obtained into a suitable form and integrate it into the decision-making process. The deployment can also go from the simple production of a report, which describes the obtained knowledge, to the establishment of an application that allows the use of the obtained model for the prediction of unknown values of an element of interest. The final cycle of the data analytics process sketched above can be schematized in the following way:

- Upstream of the analytic approach are data, and downstream refers to knowledge and then action. Therefore, please note that the importance of the data analytics process includes all the steps from recovery to deployment.
- Disciplines that support the big data analytics process.
- The phases outlined previously are supported by a range of data management and analysis disciplines, which are detailed as follows.

4.5 STATISTICS

Statistics provides the theory for testing hypotheses about various insights from data. It is intended to match the data with a predefined model whose parameters may vary. The approach generally consists of assuming that the observations follow a known distribution and then testing this hypothesis to confirm or refute it.

4.6 MACHINE LEARNING

Machine learning (ML) enables the implementation of learning agents based on data mining; ML includes several heuristic techniques. ML is a self-learning method, i.e. an artificial intelligence that allows the machine to produce estimates or forecasts whose performance will depend on the data. This allows us to say that ML is a discipline at the crossroads of big data and artificial intelligence, which presents a discipline that seeks to solve complex logical problems by "imitating" the human cognitive system.

For more clarity, let us briefly illustrate what machine learning can do with a simple case, probably closer to everyday life: *an anti-spam filter*. At first, we can imagine that the system will analyze how you will classify your incoming mails in spam. Because of this learning period, the system will deduce some criteria of classification. For example, the probability that the machine will classify a mail in spam will increase if the e-mail contains terms such as "money", "free", "win", etc. and the fact that the sender of the mail does not appear in your address book. On the other hand, the probability of ranking in spam will drop if the sender is already known and the words of the mail are more reliable. With machine learning, we move on

from imperative computing based on hypotheses to probabilistic computing based on real data. So, in addition to the importance of understanding the taxonomy of the system (“IF”, “THEN”, etc.), we need, first, the data.

4.7 DATA MINING

Data mining is a particular step in the process involves the application of specific algorithms for extracting models from data. The additional steps in the process, such as data preparation, data selection, data cleaning, incorporation of appropriate prior knowledge, and proper interpretation of the results of mining, ensures that useful knowledge is derived from the data.

Data mining and knowledge discovery combines theory and heuristics toward extracting knowledge. To this end, data cleaning, learning and visualization might be also employed. We can say that the main task of data mining is using methods to automatically extract useful information from these data and make them available to decision-makers.

4.8 TEXT MINING

Text mining is the analysis of data contained in natural language text. It refers to the technique that automates the processing of large volumes of text content to extract the key trends and to statistically identify the different topics that arise. The application of text mining techniques to solve business problems is called text analytics. Techniques of text mining are mainly used for data already available in digital format. Online text mining can be used to analyze the content of incoming emails or comments made on forums and social media.

4.9 DATABASE MANAGEMENT SYSTEMS

These systems include relational database management systems, NoSQL databases and big data databases, such as the Hadoop distributed file system, which provide the means for data persistence and management.

4.10 DATA STREAMS MANAGEMENT SYSTEMS

These systems handle transient streams, including continuous queries, while being able to handle data with very high ingestion rates, including streams featuring unpredictable arrival times and characteristics. Now that you have understood the

context of the data analytics process and you have a basic understanding of the different issues, which you need before approaching any process, it is time to get into the practical side of things. In other words, how about seeing how the model is built?

4.11 WHAT TO AVOID WHEN BUILDING A MODEL?

Remember, one of the goals of the data analytics process is to find a model that approximates the reality (the phenomenon) that by using this model we will be able to predict. A model can be seen as a mathematical function, to which we introduce input "data" that characterize the phenomenon that we want to predict, and that at the exit proposes a score or a result (output) for this phenomenon. Before creating the model, there are three essential elements to consider:

Describe: The first essential point before designing a model is the description of the phenomenon that will be modeled by determining the question to be answered. The statistical description gives a global view of trends and dominant patterns that structure, for example, the logic of purchase, contact and satisfaction. This first segmentation makes it possible to build a typology of customers and better target an offer.

Predict: A model can be used to anticipate future behavior. Big data offers unprecedented opportunities for segmentation, targeting and identifying new prospects.

Decide: Prediction tools and models provide insights that can be useful in the decision-making process. Their activities and actions will lead to better results for the company because of the "intelligence" of the model creation process. The model provides answers to questions about the anticipation of future behaviors, or the discovery of a hitherto unknown characteristic concerning a phenomenon, by detection of certain profiles or look-alikes.

Here are some characteristics to consider when choosing a model:
- interpretability;
- simplicity;
- accuracy;
- speed (testing and real-time processing);
- scalability.

Minimize the model error: A first way of representing this loss is by what is called "error", which refers to the distance between data and the prediction generated by the considered model.

Maximize the likelihood of the model: Indeed, the loss in this case is a bit hidden, but we can find mathematically that maximizing likelihood is equivalent to minimizing a loss function. The objective is to converge toward the maximum of the likelihood function of the considered phenomenon by finding from the initial observations. The loss functions are illustrative examples of the approach that is developed to build a model, because a model is a story of optimization.

A large part of the models thus lies in the optimization methods, i.e., the methods that will seek a maximum or a minimum of a determined function. Once a model is built, we want to use it with new data and new individuals. In practice, the optimization algorithms are built into the model you want to create.

4.12 SELF-ASSESSMENT QUESTIONS

Q. 1 Write data analytics process with relevant examples.

Q. 2 Describe big data analytics tools with suitable examples.

Q. 3 What is a database management system? Explain with example.

Q. 4 Explain the following:

- Data mining
- Machine Learning (ML)
- Text mining
- Data streams management systems.

4.13 ACTIVITIES

- Write the process of using Machine Learning (ML) and data mining techniques in LIS discipline?

REFERENCES

Cooper A., (2012). What is analytics? Definition and essential characteristics, CETIS Analytics Series, 1 (5). 1–10.

Kandasamy, B. P. and Benson, V. (2013). Making the most of big data: Manager's guide to business intelligence success. www.bookboon.com. http://93.174.95.29/_ads/EC133CCE54AA14A53992645E9C31BF95

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Hung Byers, A. (2011). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute.

Sedkaoui, S. (2018). Data analytics and big data. John Wiley & Sons.

Vasarhelyi, M. A., Kogan, A., & Tuttle, B. M. (2015). Big data in accounting: an overview. *Accounting Horizons*, *29*(2), 381–396.

Unit–5

DATA ANALYTICS AND MACHINE LEARNING

Compiled by: Dr. Amjid Khan

Reviewed by: 1. Dr. Pervaiz Ahmad
2. Muhammad Jawwad
3. Dr. Muhammad Arif

CONTENTS

INTRODUCTION

This unit describes the ML context and process as an important aspect of the artificial intelligence (AI) and as one of the main tools for a data scientist. It also explains descriptive analysis, predictive and prescriptive analyses with examples. At the end of the unit, self-assessment questions followed by practical activities are given to the students.

OBJECTIVES

After reading this unit, you will be able to explain:

- data analytics and machine learning.
- descriptive analysis, predictive and prescriptive analyses.
- artificial intelligence: algorithms and techniques.
- skills required for data scientist.

5.1 INTRODUCTION

This unit describes the ML context and process as an important aspect of the artificial intelligence (AI) and as one of the main tools for a data scientist.

From simple descriptive analysis to predictive and prescriptive analyses: what are the different steps?

The application of analytics can be divided into three main categories, namely descriptive, predictive and prescriptive analytics. Descriptive analytics involves using advanced techniques to locate relevant data and identify remarkable patterns to better describe and understand what is going on with the subjects in the dataset. Data mining, the computational process of discovering patterns in large datasets involving methods at the intersection of artificial intelligence (AI), ML, statistics, and database systems, is accommodated in this category.

Descriptive models can give a clear explanation as to, how and why a certain event occurred, but all of this is already perfectly in the past. So, based on the past, companies can have a clear vision of the future, on what is more important and how they can function. These appeal to predictive models that are seen as a subset of data science.

Predictive analytics use data, statistical algorithms, and ML to predict the likelihood of business trends and financial performance based on their past behaviors. They bring together several technologies and disciplines such as statistical analysis, data mining, predictive modeling, and ML technology to predict the future of businesses. With the increasing number of data, computing power and the development of AI software and simple analytical tools' uses, many companies can now use predictive analytics. Predictive analytics is the act of predicting future events and behaviors present in previously unseen data using a model built from similar past data.

- The requirements for an accurate and reliable prescriptive analytics outcome are hybrid data, integrated predictions, and prescriptions, considering side effects, adaptive ML algorithms and a clear feedback mechanism. AI and ML can be considered as a top level of data analysis.
- Cognitive computer systems constantly learn about the business and intelligently predict industry trends, consumer needs, etc. The level of cognitive applications can be defined by four main skills:
- An understanding of unstructured data;
- The ability to extract information and ideas;
- The ability to refine expertise with each interaction.

- The ability to see, hear and speak to interact with humans in a natural way. Along with mathematical, statistical and analysis methodologies, ML and big data analytics have emerged to build systems that aim at automatically extracting information from the raw data that the IT infrastructures offer.

5.2 ARTIFICIAL INTELLIGENCE: ALGORITHMS AND TECHNIQUES

Artificial intelligence found its name at the Dartmouth conference in 1956. But it began in the early 1950s with, for example, the work of Alan Turing which questioned whether we could make a computer think. He proposed a test called "the Turing test", in which a person chats through a computer and must guess if their interlocutor is a machine or a human being.

If the person cannot pass the test, then we can conclude that it is possible to operate a computer with logic algorithms like our way of thinking or even beyond. Indeed, AI has vegetated several times, especially in the 1970s and 1990s; because it has been limited for a long time by the costs and performances of the machines (speed, memory capacity, storage capacity), which have undermined the expectations that had been placed there, causing frustrations and losses of investments by industrialists.

AI had its beginnings in computer science with automated systems, recurrence and languages like "Lisp" and "Prolog". In its early days, we mainly talked about logical rules, recursion, parsing, graphs and expert systems. Most of the techniques used in AI are based on mathematical theories (advanced statistics, decision trees, Bayesian networks, neural networks, etc.) that have been known for 50 years or more. Now, the techniques used to make our machines think are many:

- fuzzy logic;
- genetic algorithms;
- data mining;
- Bayesian inference;
- smart agents;
- neural networks;
- automatic learning.

They are becoming more and more used today because of the conjunction of several factors:

- data storage costs are constantly decreasing;
- the increase of computing power;
- the explosion of the amount of information available in digital form;
- this information is largely unstructured and requires operating different from conventional methods.

ML is a data analysis technique that teaches computers what humans are naturally capable of learning from their experiences. ML's algorithms use computational methods that "learn" information directly from the data without the need to rely on a predetermined equation as a model. Algorithms adapt and become more efficient as the number of samples available for learning increases. Algorithms of ML identify natural patterns in the data that generate useful information and help to make better decisions and predictions. They are used daily to make critical decisions in medical diagnostics, stock trading, energy load forecasts and more. For example, websites take advantage of ML to process millions of options to recommend songs to listen to or movies to watch. Retailers use it to understand the buying behaviors of consumers.

With the rise of big data, ML has emerged as one of the best problem-solving techniques in some areas, including the following:
***Finance*:** banks and insurances use ML to discover important information within the data and to prevent fraud. Also, for credit evaluation and algorithmic trading.

***Health*:** ML is being used more and more in the healthcare industry, particularly with the rise of connected objects and other sensors that make it possible to use the data to access a patient's health data in real time. This technology can also help medical experts analyze data to identify alarming trends to improve diagnostics and treatments.

***Marketing*:** websites that recommend products based on a user's previous purchases use ML to analyze the customer's purchase history and offer products that might interest them. The ability to collect, analyze and use data to personalize the shopping experience represents the future of retail.

***Image processing and computer vision*:** they are used for facial recognition, motion detection and object detection, or recognition of friends in photo albums or via search engines. Also, voice recognition (during a phone call) to dictate a text to your smartphone or to recognize a song on the radio (Shazam, Sound Hound applications, etc.).

***Biology*:** ML is used for tumor detection, drug discovery and DNA sequencing.

Energy production: it is used for forecasting prices and charges. ML can also find new sources of energy, analyze minerals in the soil, or predict sensor failures in refineries. This technology makes oil distribution more efficient and economical;
***Automotive, Aerospatiale and industrial production*:** they are used for predictive maintenance or for the automatic control of our cars, etc.;

***Natural language processing*:** it is used for speech recognition applications;
Art also does not escape: it is used for making music, poetry or even paintings.

5.3 ML: WHAT IS IT?

ML refers to all the approaches that give computers the ability to learn autonomously. These approaches, which overcome strictly static programs for their ability to predict and make decisions based on the data input, were used for the first time in 1952 by Arthur Samuel, one of the pioneers of AI, for a game of checkers. Samul defines ML as the field of study aimed at giving a machine the ability to learn without being explicitly programmed. Tom Mitchell of Carnegie Mellon University proposed a more precise definition:

A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E". The basic idea of ML is that a computer can automatically learn from experience. Using collected data, a ML algorithm finds the relations between different properties of the data. The resulting model can predict one of the properties of future data based on properties.

Although ML applications vary, its general function is similar throughout its applications. The computer analyzes a large amount of data, and finds patterns hidden in it. These patterns are mathematical in nature, and they can easily be defined and processed by a machine. Though, it is currently boosted by new technologies and new uses, ML algorithms have been widely adopted in different fields such as business, computer science and so on. To learn and grow, computers need data to analyze and train. In fact, big data is the essence of ML, and ML is the technology that makes full use of the big data potential. For the analysis of such amounts of data, ML is much more efficient than traditional methods in terms of accuracy and speed. For example, based on data associated with a transaction, ML can detect potential fraud in a millisecond. Thus, this method is much more efficient than traditional methods for analyzing transactional data or data from social networks or CRM platforms.

Their extensive application is due to their ability to automatically extract information from the data. In the context of IT, ML techniques have been used to solve many problems related to anomaly detection, patterns discovery, profiling, etc. For example, by analyzing website content, search engines can define which words and phrases are the most important in defining a certain web page, and they can use this information to return the most relevant results for a given phrase search.

ML improves diagnostics, predicts better outcomes and is revolutionizing personalized care. The basic idea of any ML process is to train the model, based on some algorithm, to perform a certain task: classification, clustering regression, etc. ML can be used to reveal a hidden class structure in unstructured data, or it can be used to find dependencies in a structured data to make predictions. So, as part of the job of the ML process, it is supposed that a relationship exists between available data,

and it is the role of algorithms to reveal it. To reveal this relationship, which can lead you to extract value; you must adopt the algorithm that will allow you to do this.

5.4 HOW DOES ML WORK?

In the AI field, which aims to make machines perform tasks that normally require human intelligence, ML is currently the dominant trend. With ML, the computer performs tasks for which it has not been explicitly programmed to complete by producing models itself and sometimes even changing them from new data. ML algorithms use large amounts of data. They are getting closer to data mining or BI (business intelligence). However, data mining is limited to making data intelligible by presenting them analytically and synthetically. ML goes further by producing rules or models that can explain the data, thus potentially predicting new data (predictive analytics), or even ultimately making decisions based on new data and the established model It should be noted, before speaking about the ML process, that ML relies on two types of techniques: supervised learning, which involves training a model on known input and output data, so that it can predict future outcomes, and unsupervised learning, which identifies hidden models or intrinsic structures in the input data. Working with ML algorithms means that a whole workflow is taking place. It will include the following:

- **Definition the business need (problem statement) and its formalization**
  Identify the main learning problem by considering what is observed and the answer you want the model to predict. Defining the problem to solve helps to clarify ideas. So, the first step is to imagine a path between the initial data and the value to be predicted. At this stage, we can describe the problem in an *informal* way. In other words, we mean to formulate the problem in a precise and concise sentence.

- **Collection and preparation of the useful data that will be used to meet this need**
  This step consists of collecting and registering the type of data useful for solving the problem. Collect, clean, and prepare data so that they can be consumed by ML model training algorithms.

- **Test the performance of the obtained model**
  Several types of problems are solved by the ML algorithms, but it is necessary to choose the algorithm that allows us to better model the problem, including classification, regression, and clustering. With each type of problem, one can have several candidate algorithms to solve it. The factors that can come into play in choosing the right algorithm can be many, including the number of features, the amount of data we have, etc. After rolling out its algorithm on its training set and making predictions with the test set, it is time to evaluate its performance. There are some standard ways to do this depending on the

types of problems. For example, for regression a model can be considered good by an indicator.

- **Optimization and production start**
  The results of an ML problem are often complex to interpret. Whatever the method adopted in the analysis process, the following points should be answered: ***the context*****:** raising the context of the problem and the reasons for its resolution.

  the problem**:** concisely describe the problem we are trying to solve;

  the solution**:** describe the solution provided in terms of architecture, how to exploit the solution, etc.

  limitations**:** if the solution is not universal or has limitations, it is better to list them. This gives a solution credibility and can open paths to new areas of improvement.

  conclusion**:** quickly revisit the description of the problem as well as the solution and the benefits derived from it.

5.5 SKILLS REQUIRED FOR DATA SCIENTIST

The following skills are required for a data scientist.

- **Probability and Statistics**
  Data Science is about using capital processes, algorithms, or systems to extract knowledge, insights, and make informed decisions from data. In that case, making inferences, estimating, or predicting form an important part of Data Science. Probability with the help of statistical methods helps make estimates for further analysis. Statistics is mostly dependent on the theory of probability. Putting it simply, both are intertwined. What can you do with Probability and Statistics for Data Science? Explore and understand more about the data Identify the underlying relationships or dependencies that may exist between two variables Predict future trend or forecast adrift based on the previous data trends Determine patterns or motive of the data Uncover anomalies in data Especially for data-driven companies where stakeholders depend on data for decision making and design/evaluation of data models, probability and statistics are integral to Data Science.

- **Multivariate Calculus and Linear Algebra**
  Most machine learning, invariably data science models, are built with several predictors or unknown variables. A knowledge of multivariate calculus is significant for building a machine learning model.

- **Programming, Packages and Softwares**
  Programming Skills for Data Science brings together all the fundamental skills needed to transform raw data into actionable insights. While there is no specific rule about the selection of programming language, Python and R are the most favored ones for data science.

- **Data Wrangling**
  Data Wrangling is the process where you prepare your data for further analysis; transforming and mapping raw data from one form to another to prep up the data for insights. For data wrangling, you basically acquire data, combine relevant fields, and then cleanse the data.

- **Database Management**
  For me, data scientists are different people, master of all jacks. They must know math, statistics, programming, data management, visualization, and what not to be a "full-stack" data scientist.

  Database Management quintessentially consists of a group of programs that can edit, index, and manipulate the database. The DBMS accepts a request made for data from an application and instructs the OS to provide specific required data. In large systems, a DBMS helps users to store and retrieve data at any given point of time.

- **Data Visualization**
  It is a graphical representation of the findings from the data under consideration. Visualizations effectively communicating and lead the exploration to the conclusion. It gives you the power to craft a story from data and create a comprehensive presentation. Data Visualization is one of the more essential skills because it is not just about representing the results, but also understands and learns the data and its vulnerability. Histograms, Bar charts, Pie charts, Scatter plots, Line plots, Time series, Relationship maps, Heat maps, Geo Maps, 3-D Plots, and a long list of visualizations you can use for your data.

- **Machine Learning / Deep Learning**
  ML is a subset of the Data Science ecosystem, just like Statistics or Probability that contributes to the modeling of data and obtaining results.

  Machine Learning for Data Science includes algorithms that are central to ML; K-nearest neighbors, Random Forests, Naive Bayes, Regression Models. PyTorch, TensorFlow, Keras also find its usability in Machine Learning for Data Science.

- **Cloud Computing**
  The practice of data science often includes the use of cloud computing products and services to help data professionals access the resources needed to manage and process data. An everyday role of a Data Scientist generally

includes analyzing and visualizing data that are stored in the cloud. Familiar with the fact that data science includes interaction with large volumes of data, given the size and the availability of tools and platforms, understanding the concept of cloud and cloud computing is not just a pertinent but critical skill for a data scientist.

- **Microsoft Excel**
  MS Excel is one of the best and most popular tools to work with data. What can you do with Excel for Data Science?
  - Naming and creating ranges
  - Filer, sort, merge, trim data
  - Create Pivot tables and charts

  Visual Basic for Applications (VBA), an MS Excel superpower programming language of Excel which allows you to run loops, macros, etc. Clean data: remove duplicate values, change references between absolute, mixed, and relative Look-up required data among thousands of records.

- **DevOps**
  DevOps is a set of methods that combines software development and IT operations that aims to shorten the development life cycle and provide uninterrupted delivery with high software quality. DevOps teams closely work with the development teams to manage the lifecycle of applications effectively. Data transformation demands close collaboration of data science teams with DevOps. What can be done with DevOps for Data Science?
  - Provision, configure, scale, and manage data clusters.
  - Manage information infrastructure by continuous integration, deployment, and monitoring of data.
  - Create scripts to automate the provisioning and configuration of the foundation for a variety of environments.

- **Excellent communication skills are also needed**. Here is a summary of the various types of communication skills that a data scientist should have a command over.
  - Presentation skills.
  - Storytelling skills.
  - Business insight skills.
  - Writing / publishing skills.
  - Social media skills.
  - Listening skills.
  - Stop and Thinking skills.

5.6 SELF-ASSESSMENT QUESTIONS

Q. 1 Write a comprehensive note on data analytics and machine learning.

Q. 2 Describe descriptive analysis, predictive and prescriptive analyses with relevant examples.

Q. 3 What is artificial intelligence? Explain with examples.

Q. 4 What skills are required for data scientist? Discuss.

Q. 5 How does ML work? Explain with examples.

5.7 ACTIVITIES

Now you know how to identify the raw material for data analysis and ML processing and you have an idea about what you can and want to do with it. As an activity for you, identify the raw data in library science context for analysis and ML processing and how to interpret it.

REFERENCES

Cooper A., (2012). What is analytics? Definition and essential characteristics, CETIS Analytics Series, 1 (5). 1–10.

Kandasamy, B. P. & Benson, V. (2013). Making the most of big data: Manager's guide to business intelligence success. www.bookboon.com. http://93.174.95.29/_ads/EC133CCE54AA14A53992645E9C31BF95

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Hung Byers, A. (2011). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute.

Sedkaoui, S. (2018). Data analytics and big data. John Wiley & Sons.

Vasarhelyi, M. A., Kogan, A., & Tuttle, B. M. (2015). Big data in accounting: an overview. *Accounting Horizons*, *29*(2), 381–396.

Unit–6

SUPERVISED VERSUS UNSUPERVISED ALGORITHMS: A GUIDE

Compiled by: Dr. Amjid Khan

Reviewed by: 1. Dr. Pervaiz Ahmad
2. Muhammad Jawwad
3. Dr. Muhammad Arif

CONTENTS

INTRODUCTION

The aim of ML is to train algorithms so that they can learn to make predictions on a large amount of data. Depending on the type of input data, machine learning algorithms can be divided into supervised and unsupervised learning. This unit primarily focuses on the major concepts to develop an effective roadmap for implementing a supervised and unsupervised ML algorithm. It also describes different techniques for advanced supervised and unsupervised algorithms, such as clustering, classifications, and regression models. At the end of the unit, self-assessment questions followed by practical activities are given to the students.

OBJECTIVES

After reading this unit, you will be able to understand:

- important concept to develop an effective roadmap for implementing a supervised and unsupervised ml algorithm.
- how to transform your business objectives into a data analysis process using the ml process?
- different techniques for advanced supervised and unsupervised algorithms, such as clustering, classifications, and regression models.

6.1 INTRODUCTION

The goal of ML is to train algorithms so that they can learn to make predictions on a large amount of data. Depending on the type of input data, machine learning algorithms can be divided into supervised and unsupervised learning. Supervised and unsupervised learning comes from data mining, which aims to extract knowledge from databases.

6.2 SUPERVISED LEARNING: PREDICT, PREDICT AND PREDICT!

A *supervised learning* task is called "classification" if the outputs are discrete or "Regression" or if the outputs are continuous. In supervised learning, input data comes with a known class Structure. If an algorithm is given a set of inputs (features vectors):

- $\{x_1, x_2, \ldots x_n\}$, and a set of corresponding outputs (labels):
- $\{y_1, y_2, \ldots yn\}$ then the goal of the algorithm is to learn to produce the correct output given a new input; this is a supervised learning task.

The algorithm is usually tasked with creating a model that can predict one of the properties by using other properties. After a model is created, it is used to process data that has the same class structure as input data. Inputs can be vectors of different types of objects, integer numbers, real numbers, strings, or more complex objects. Outputs take values, each representing a unique state. For example, an algorithm may be given several vectors representing numerically external features of a person, such as sex, age, income, etc., and corresponding outputs that take one value from the set "male, female". Supervised algorithm can also be applied in the detection of spam from your mail as well as in the forecast scores and risks associated with insurance. But how does it work? Let us take a simple example: We will propose a series of pictures, knowing that the target that we want will be the "category". The objective is then to classify, by means of classification methods, the group of membership of each picture according to its similarity to other pictures.

In supervised learning, you will retrieve annotated data from their outputs to train the model, i.e., you have already associated them with a label or a target class and you want that the algorithm be able to predict it for new non-annotated data once trained. So, the system learns to classify according to a predetermined classification model and known examples.

Supervised learning is divided into the following two parts:

- **The first is to determine a tagged data model.**
  The second consists of predicting the label of a new datum, knowing the previously learned model. Supervised learning will be applied when the goal is to predict a value or belonging.

- **Unsupervised learning: go to profiles search!**
  If an algorithm is only given a set of inputs: $\{X_1, X_2, \ldots X_n\}$, and no outputs, this is an *unsupervised learning* task. Unlike supervised learning, which attempts to find a model from labeled data: (X) → Y, unsupervised learning takes only untagged data (no variable to predict Y). In unsupervised learning, input data do not have a known class structure, and the task of the algorithm is to reveal a structure in the data.

  An unsupervised learning algorithm will find patterns or structuring in the data. In unsupervised learning, there is no initial labeling of data. Here, the goal is to find some pattern in the set of unsorted data, instead of predicting some value. Unsupervised methods usually generate too many false alerts, so it is often a good idea to combine both supervised and unsupervised methods.

  Unsupervised learning can be thought of as finding patterns in the data and beyond what would be considered as pure unstructured noise. In unsupervised learning, input data are not annotated. So, *how can this work?* Good question. The algorithm must discover by itself the pattern according to the data. The algorithm applies in this case to finding only the similarities and distinctions within these data, and then grouping together those that share common characteristics.

  Clustering algorithms fall into the unsupervised learning category. They make it possible to group together similar data. Find the hidden patterns in the unlabeled data and separate it into clusters according to similarity. An example can be the discovery of different customer groups inside the customer base of the online shop. For example, in the case where Amazon receives a new purchase proposal from you (as a new user), Amazon users are divided into groups and, according to your purchase choice, you will be associated with a group of clients who have purchases close to yours. It is just about bringing clients into groups that are not predefined. Relating this back to our previous example of the categorized pictures, if we inject thousands more photos, similar pictures would be automatically grouped within the same category.

6.3 REGRESSION VERSUS CLASSIFICATION

Supervised learning assumes the availability of labeled samples, i.e. Observations annotated with their output, which can be used to train a learner. In the training set, we can distinguish between input features and an output variable that is assumed to be dependent on the inputs. The output, or response variable, defines the class of observations, and the input features are the set of variables that have some influence on the output and are used to predict the value of the response variable.

Another distinction that will help you in the choice of a machine learning algorithm is the type of output expected from our program: is it a continuous value (a number) or a discrete value (a category)? The first case is called a regression, and the second is called a classification. The first assumes a categorical output, while the latter a continuous one. So, depending on the type of output variable we can distinguish between two types of supervised task: Classification and regression.

For example, if you want to determine the cost per click of a web advertisement, you apply a regression. If you trying to determine if a photo is of an apple or a banana, you apply a classification analysis. The regression/classification distinction is about supervised algorithms. It distinguishes two types of output values that can be sought to be processed.

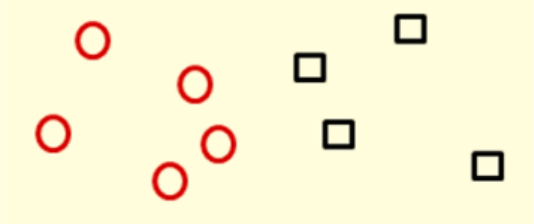
- **Regression**
  Regression method takes a finite set of relations between dependent variables and independent variables and creates a continuous function to generalize these relations. Regression predicts the value based on previous observations, i.e., values of the samples from the training set. For example, let us suppose that you want to predict the income of clients and their prospects based on data such as their socio-professional category, age, gender, occupation, address and so on. You can collect many observations by conducting a survey on a panel of clients. Then, some of these observations will be used to generate a model that can predict this income. The linear regression assumes that the data come from a phenomenon that has the shape of a straight line, i.e., there is a linear relationship between the input or the observations and the output, which take the form of predictions.

- **Classification**
  In contrast with regression problems, when the explained variable is a value in a finite set, it is referred to as a supervised classification problem. This amounts to assigning a label to each observation. Classification is a particular supervised learning task. This is called tags assigned to the input values. This is the case of: “true/false” or “passed/failed”. This method can also be used,

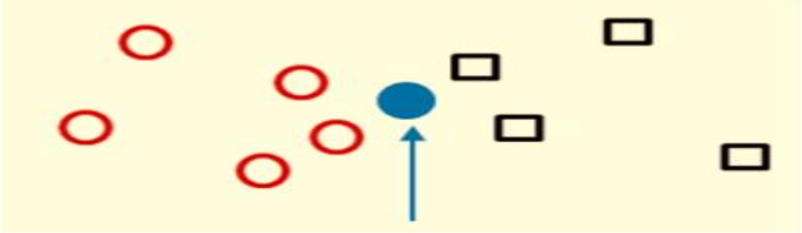
for example, in health risk analysis. A patient's vital statistics, health history, activity levels and demographics can be cross referenced to score (the level of a risk) and assess the likelihood of illness. When the set of possible values of a classification exceeds two elements, we speak of multiclass classification. Figure 7.6 illustrates both types of classification. Among the classification algorithms, we find the K-nearest neighbors (kNN), logistic regression, Support Vector Machine (SVM), Naïve Bayes, decision tree, Random Forest, and neural networks.

- **K-Nearest Neighbors**

  The KNN is an algorithm that can be used for both classification and regression. The principle of this model consists of choosing the k data closest to the studied point to predict its value. In classification or regression, the input will consist of the k closest training examples in a space. To understand the functioning of this algorithm, we will take a small visual example. Below, we will show a training dataset, with two classes, circle and square. So, the input is bidimensional, and the target is to classify the data by shape.

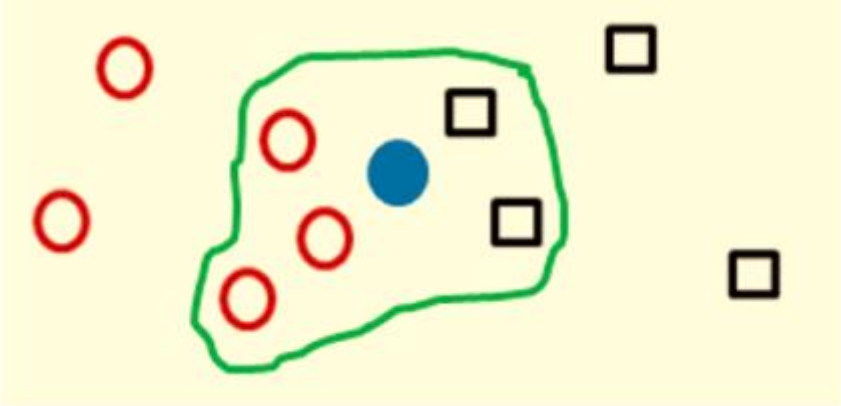


Now, if we have a new entry object whose class we want to predict, how could we do it?



The dark circle is a new entry

Well, we will simply look at the k closest neighbors to this point and see which class constitutes most of these points in order to deduce the class of the new entry.

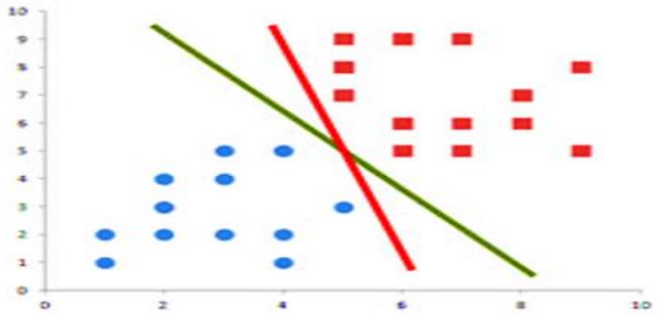


For example, if we use the 5-NN, we can predict that the new entry belongs to the circle class since it has three circles and two squares in its entourage. So, the principle of this algorithm is to classify a dataset in one of the categories by calculating the distance between it and each point of the training set. We choose the first k elements in order of distances, and therefore choose the dominant label among the k elements, which represents the category of the dataset element.

- **Logistic Regression**
  This is a statistical method for performing binary classifications. It takes qualitative and/or ordinal predictors as input and measures the probability of the output value using the sigmoid function. We can perform the multiclass classification (for example, classify a picture into three possibilities such as fruits, legumes, and roses).

- **Support Vector Machine**
  SVM is also a binary classification algorithm. As shown in below, blue represents a class (non-spam mail for example) and red represents a spam. After tagging some words and concepts, the "signature" of the message can be injected into a classification algorithm to determine whether it is a spam. Logistic regression can separate these two classes by defining the line in red. This method will opt to separate the two classes by the green line.

  Without going into details, and for mathematical considerations, the SVM will choose the clearest separation possible between the two classes (like the green line). Therefore, it is also called large margins classifier.



Example of SVM

- **Naïve Bayes**
  Naïve Bayes is an intuitive classifier to understand. It assumes a strong (naïve) assumption. Indeed, it assumes that the variables are independent of each other. This simplifies the calculation of probabilities. Generally, Naïve Bayes is used for text classifications (based on the number of word occurrences). Naïve Bayes classification is a machine learning method relying on Bayes' theorem: It can be used for both binary and multiclass classification problems.

  The main point relies on the idea of treating each feature independently. The Naïve Bayes method evaluates the probability of each feature independently, regardless of any correlations, and makes the prediction based on Bayes' theorem. That is why this method is called "naïve"; in real-world problems, features often have some level of correlation between each other. The advantages of using this method include its simplicity and ease of understanding. In addition, it performs well on the data sets with irrelevant features, since the probabilities contributing to the output are low. Therefore, they are not considered when making predictions. Moreover, this algorithm usually results in good performance in terms of consumed resources, since it only needs to calculate the probabilities of the features and classes; there is no need to find any coefficients like in other algorithms. Its main drawback is that each feature is treated independently, although in most cases this cannot be true.

- **Decision Tree**
  Another classification method is that of the decision tree. Decision trees are graph structures where each potential decision creates a new node, resulting in a tree-like graph. The decision tree is an algorithm based on a graph model (the trees) to define the final decision. Each node has a condition, and the connections are based on this condition (true/false or yes/no or pass/fail).

The further we descend into the tree, the more we combine the conditions. A decision tree is built using a machine learning algorithm. Going from a set of predefined classes, the algorithm searches iteratively for the most different variables in the classified entities. Once this is identified, and the decision rules are determined, the dataset is segmented into several groups according to these rules. Data analysis is performed recursively on each subset until all key classification rules are identified.

- **Random Forest**
  Random Forest is one of the most popular machine learning algorithms. It requires almost no data preparation and modeling but usually produces accurate results. Random Forests are based on the decision trees described previously. More specifically, Random Forests are collections of decision trees, producing better prediction accuracy. That is why it is called a "forest" – it is basically a set of decision trees.

- **Neural Networks**
  Neural networks are inspired by the neurons of the human nervous system. They allow us to find complex patterns in the data. These neural networks learn a specific task based on the training data. Neural networks consist of nodes. In these networks, we find the input third (input layer) that will receive the input data. The input layer will then propagate the data to hidden layers. Finally, the third party (output layer) produces the classification result. Each third of the neural network is a set of interconnections of the nodes of one third with those of the other thirds.

  In these networks, the learning phase aims to converge the data parameters into an optimal classification. They require a lot of learning data and are not suitable for all problems, especially if the number of input parameters is too low. In this case, the term *deep learning* refers to networks of juxtaposed neurons or consists of several layers. It draws upon, among other things, the latest advances in neuroscience and communication models of our nervous system. Some also associate it with modeling that provides a higher level of data abstraction to produce better predictions.

  Deep learning is particularly effective on the processing of images, sound and video. It is found in the fields of health, robotics, computer vision, etc. Each of the algorithms cited above has its own mathematical and statistical properties. Depending on the training set and our features, we will opt for one or the other of these algorithms.

- **Clustering Gathers Data**
  ML is undoubtedly one of the major assets in understanding the challenges of society of today and tomorrow. Among the different components that make up this discipline, we will focus on one of the sub-domains of application that characterize it: “clustering”. This field covers diverse and varied subjects and makes it possible to study the associated problems according to different perspectives. Indeed, we speak here of algorithmic objectivity.

6.4 CLUSTERING AIMS AND ITS APPLICATION IN DIFFERENT DISCIPLINES

The objective of clustering is to divide a set of objects, represented by inputs: into a set of disjointed clusters: which contain objects like each other in some sense.

Clustering aims to determine a segmentation of the studied population without *a priori* knowledge on the number of classes or “clusters”, and to interpret *posteriori* the clusters thus created. Here, humanity does not need to assist the machine in its different discovery typologies, since no target variable is provided to the algorithm during its learning phase.

Clustering algorithms are most often used for exploratory data analysis. This is, for example, to identify customers with similar behaviors (market segmentation), users who have similar uses, communities in social networks, etc. The goal is to place the entities in a single large pool and form smaller groups that share similar characteristics; find the hidden patterns in the unlabeled data and separate it into clusters according to similarity.

The efficiency of applying a clustering algorithm can allow a significant increase in the turnover of an e-commerce site such as Amazon, for example. Also, if we provide a set of pictures of animals without specifying what animals they are, then the algorithm will group together, for example, the pictures of dogs, of tigers and so on.

A cable TV that wants to determine the demographic distribution of network viewers can do so by creating clusters from available subscriber data and what they are watching. Another example can refer to the discovery of different customer groups inside the customer base of an online shop. Even a restaurant chain can group its customers according to the menus chosen by geographic location, and then modify its menus accordingly.

We can also mention the biomedical field as one of the extended fields of application of this algorithm, though, among other things, the grouping of differential genes according to their expression profile in a biological phenomenon

over time. Also, for most music lovers, these algorithms can also be used, as already mentioned, for recommendation to distribute different music in clusters and to propose a song "similar" to that to which we just listened.

Some people on the web are planning to apply these methods to *Game of Thrones* characters to detect typologies of individuals, and who knows, perhaps scientifically determine who will be the real contenders for the iron throne.

We can think more specifically of the detection of fraud, whether in public transport, as part of a complementary health reimbursement or regarding energy consumption. In general, clustering algorithms examine a defined number of data characteristics and map each data entity to a corresponding point in a dimensional chart. The algorithms then seek to group the elements according to their relative proximity to each other in the graph.

6.5 PRINCIPLE OF CLUSTERING ALGORITHMS

When it comes to the non-labeled data, the ML algorithm will group these data by similarity. Since we are talking about similarity, we must also talk about "clustering". This is a family of unsupervised algorithms. Clustering algorithms fall into the unsupervised learning category. These make it possible to group together data that are similar. This algorithm is an unsupervised learning task.

Clustering consists of grouping the data into homogeneous groups called classes or clusters, so that the elements within the same class are similar, and the elements belonging to two different classes are different. It is therefore necessary to define a measure of similarity between two elements of the data: the distance. Each element can be defined by the values of its attributes, or what we call, from a mathematical point of view, a vector. So, clustering refers to the methods of automatically grouping data that are most like one set of "clusters". A set of unsupervised algorithms can accomplish this task. They therefore automatically measure the similarity between the different data. Clustering algorithms therefore depend strongly on how we define this notion of similarity, which is often specific to the application domain. The principle of the algorithm consists of assigning classes according to:

- Minimizing the distance between the elements of the same cluster (intra-class distance);
- Maximizing the distance between each cluster (interclass distance).
- Partitioning your data by using the K-means algorithm K-means is a type of clustering algorithm that is commonly used.

This algorithm divides a set of data entities into groups, where k is the number of groups created. The algorithms refine the assignment of entities to different clusters by iteratively calculating the average midpoint or centroid of each cluster. Let us

suppose that you are looking to launch an advertising campaign and that you want to send a different advertising message depending on the target audience. First, you need to group the target population into groups. Individuals in each group will have a degree of similarity (age, gender, salary, etc.). That is what the K-means algorithm will do. The centroids become the focal points of the iterations, which refine their locations in the plot and reassign the data entities to fit the new locations. An algorithm is repeated until the groupings are optimized and the centroids do not move anymore. The algorithm thus works as follows:

Initialization: since the number of classes كis imposed, choose points randomly to initially constitute the representatives of each class;
Then, for each point:

- calculate the distances between this point and the classes' representatives: we begin by randomly choosing *K* centroids from our observations. Each point is then associated with the centroid of which it is closest; at this point assign the class from which its distance is minimal, thus forming clusters;
- Update the representatives of each class: we can now recalculate the centroid of each cluster (its center of gravity). Repeat the operation until the algorithm converges. The illustration of the *K*-means algorithm execution result will help you to understand how it works.

Data with one to three dimensions can be represented graphically. The others can only be apprehended mathematically. Here is a simplified representation of the iterative process that the algorithm will perform on two-dimensional data. So, this algorithm solves the following problem: "given points and an integer *K*, the problem is to divide the points in *K* partitions so as to minimize a certain function".

Its speed compared to other algorithms makes it the most-used clustering algorithm. In practice, كdoes not correspond to the number of clusters that the algorithm will have to find and in which the elements will be stored, but rather to the number of centers (the central point of the cluster). It seems as if that is pretty much the same thing, but that allows us to go further. Indeed, a cluster can be represented by a circle composed of a central point and a radius.

An algorithm indicates how to combine and associate the data to obtain a response. So, to carry out a process of data analytics using ML algorithms, it is best to consider an algorithm as a recipe, and data as ingredients, while the machine is like a mixer that supports many of the difficult tasks of the algorithm.

6.6 SELF-ASSESSMENT QUESTIONS

Q. 1 Describe supervised versus unsupervised algorithms with relevant examples.

Q. 2 Differentiate between regressions versus classification.

Q. 3 What is unsupervised learning? Discuss with examples.

Q. 4 What is clustering? Explain the basic principles of clustering with suitable examples.

Q. 5 Write short notes on the following:

- Predict
- Regression analysis
- Regression Models

6.7 ACTIVITIES

- As a student of information science, you should transform your library objectives into a data analysis process using the ML process.
- Also, discover different techniques for advanced supervised and unsupervised algorithms, such as clustering, classifications, and regression models.

REFERENCES

Cooper A., (2012). What is analytics? Definition and essential characteristics, CETIS Analytics Series, 1 (5). 1–10.

Kandasamy, B. P. and Benson, V. (2013). Making the most of big data: Manager's guide to business intelligence success. www.bookboon.com.http://93.174.95.29/_ads/EC133CCE54AA14A53992645E9C31BF95

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., and Hung Byers, A. (2011). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute.

Sedkaoui, S. (2018). Data analytics and big data. John Wiley & Sons.

Vasarhelyi, M. A., Kogan, A., and Tuttle, B. M. (2015). Big data in accounting: an overview. *Accounting Horizons*, *29*(2), 381–396.

Unit–7

DATA-DRIVEN COLLECTIONS MANAGEMENT; USING DATA TO DEMONSTRATE LIBRARY IMPACT AND VALUE

Compiled by: Dr. Amjid Khan

Reviewed by: 1. Dr. Pervaiz Ahmad
2. Muhammad Jawwad
3. Dr. Muhammad Arif

CONTENTS

INTRODUCTION

Data-driven collections management is not a new concept for libraries. Today's collections librarians are all too familiar with the harsh decisions involved in building relevant and quality collections. From reading lists to user recommendations, through to library management systems and the careful analysis of data in spreadsheets and other systems, using data to inform collections management and policy is a key part of curating a library collection. This unit covers all topics related to library collections, patron-driven acquisition, gamifying collections, and managing national' collection, etc. At the end of the unit, self-assessment questions followed by practical activities are given to the students.

OBJECTIVES

After reading this unit, you will be able to explain:

- the library collections.
- managing the local collection
- patron-driven acquisition
- gamifying collections
- managing the 'national' collection
- national/regional shared collections strategies and services
- using data to demonstrate library impact and value

7.1 INTRODUCTION

Data-driven collections management is not a new concept for libraries. From reading lists to user recommendations, through to library management systems and the careful analysis of data in spreadsheets and other systems, using data to inform collections management and policy is a key part of curating a library collection. It is therefore of little surprise that libraries and librarians have, over the past few years, become increasingly interested in exploring more sophisticated and joined-up ways of taking advantage of library transaction and management data to help drive more informed and open approaches to decision making on collection management.

You can read the following two case studies which explores some of the most recent and innovative examples of how libraries are refining their collection-management processes by creating tools and applications that can utilize data to make more informed decisions about a wide range of collection-management decisions.

Case Study 2.1 From Harvard University Library, explores the creation of a library analytics toolkit and dashboard with a primary focus on collection data. Here the data and visualizations are aimed both at supporting the librarians in their everyday decision making and at enabling users to see how collections have changed over time.

Case Study 2.2: Describes the work of the Copac Collections Management (CCM) tool and the development of a prototype shared collections management service for UK academic libraries – a service that will enable both local holdings analysis and comparison across other Copac research and specialist libraries in the UK:

> **Case Study 2.1:** Dulin, K. and Spina, C., *Building an analytics toolkit at the Harvard Library* (Harvard University), p. 28.
>
> **Case Study 2.2:** Cousins, S. and Massam, D., *Collection management analytics: The Copac Collection Management tools project* (Mimas, University of Manchester), p. 35.

7.2 THE COLLECTIONS TURN

Libraries increasingly find themselves in a double bind. While budgets are reduced or remain static, user demand for access to the library, its services and content continues to grow. Further compounding this situation are the changing expectations and demands on space from users, meaning that large physical

collections need to be rethought and space must be reconfigured to meet the changing demands of users.

At the same time, e-books and digital monographs present libraries with their own challenges. In the public sector especially, the delivery of e-books continues to be problematic, as the technical and legal issues blight the user experience and compromise the library's ability to deliver them to users. And, while students and researchers have largely embraced digital access to journal articles, the same is not true for the scholarly monograph.

Research suggests that academics require the hard-copy text for long-form reading and in-depth research, especially within the humanities, social sciences, and mathematics. For a long time in academic libraries, it has been journals that have dominated discussion of academic print resources, while the book (or monograph) has been a somewhat neglected part of the collection's management debate. This inattention is giving way to a sustained focus by libraries, researchers, funders, and systems vendors, as there is an increasing realization that library print collections must be carefully and skillfully managed, space must be re-engineered for more social and collaborative uses and physical books must increasingly be seen as part of a larger collection, whether institutional/ organizational, regional, or even national.

7.3 MANAGING THE LOCAL COLLECTION

The collections turn that we are beginning to describe begins and ends with the local collection. Libraries must ensure that their collections continue to provide users with access to relevant resources and to respond to changing demands. The collections turn is primarily a recognition that collections are not built to be great collections in themselves but are there to serve their users:

– **to connect them with the content they need**.
  Libraries are responding by introducing several innovative approaches to local collections management that are reducing the 'friction' between the user and the collection, in terms of both more 'formal' collections management and acquisition approaches and more playful or 'informal' approaches to collections data.

7.4 PATRON-DRIVEN ACQUISITION

Patron-driven acquisition (PDA) is an acquisitions approach which is driven by the user. Rather than the library purchasing materials on the user's behalf, and with a 'just in case' approach, PDA enables the user to trigger the purchase of material through the action of clicking on a catalogue link or similar. This 'just in time' approach has several distinct advantages for the library given us under:

***Cost effectiveness*:** Only books or material that users want to access are purchased. In theory, the library isn't purchasing anything until a user clicks on a link to a book that they want to read.

***Increased usage and circulation*:** Not only are it the case that the material is purchased at the point of need (ensuring that the content is used at least once), but it also tends to be that PDA material has a higher circulation in general. What one user thinks is worth reading tends to agree with what others think is worth reading too.

***Collection balance*:** Despite the fears, research suggests that PDA and other forms of user-driven acquisition help the development of a balanced collection. Approaches like PDA enable users to directly affect the collections they use. But the circulation and management data used by libraries to make collections-management decisions, can also have a role in more playful approaches to engaging users with the collections.

7.5 GAMIFYING COLLECTIONS

Gamification 'involves applying game design thinking to non-game applications to make them more fun and engaging' (http://gamification.org). While this is still a very new way to engage library users, several projects and companies are beginning to explore the potential for gamifying the library experience. One of the best known in the UK is Library Game (http://librarygame. co.uk). Library Game uses library systems data – not necessarily the kind of data we might associate with an innovative, user-centered game – to collect participating users' activity and transaction data. As a Library Game player borrows or returns a book that data is used by the game to provide a social element to the circulation process, showing what other users are borrowing, etc., as well as awarding points for different activities and moving the user up (or down) the scoreboard. Users can also leave reviews of books, create friends' lists, and see their borrowing history. Importantly, Library Game also provides that data back to the library so that it can analyze behaviors and borrowing patterns to inform both its services and collections, utilizing a depth of data not previously available.

Libraries are also utilizing existing applications and services, such as SCVNGR (www.scvngr.com), to encourage students and users to undertake tasks within the library; specifically, they are using them for library inductions. This engagement is both about making the experience more fun and about using that engagement to generate intelligence and data that can be fed back to the services and collections to continually improve them. These approaches create a positive feedback loop where gradually more in-depth and richer data is generated that is used to further improve and refine services and collections. These new and emerging approaches to local collections development, strategy and management become even more critical as the data begins to inform collections-management decision making beyond the local institution and library, at a sector, regional and even national level.

7.6 MANAGING THE 'NATIONAL' COLLECTION

Several interesting developments are taking place in libraries to help transform local collections management and strategies. And, as library and collections data become more accessible, better curated, timelier, and more accurate it begins to open possibilities beyond the local library and enables decisions to be made on a regional or national level. This regional or national organization of collections describes the second part of the collections turn: the local collection is situated within and contributes to a broader regional, sector or national collection (or collections).

This increasing interest in the challenges that the transition from print to digital monographs presents to libraries and institutions is something that is being explored as part of the National Monograph Strategy (NMS) in the UK. The project is exploring the potential for a national approach to the creation, collection, preservation, and digitization of scholarly monographs. The project has outlined eight high-level ideas to address some of the challenges presented by the scholarly monograph, including a national monograph knowledge base, which would provide a comprehensive and open bibliographic and holdings database enabling the development of new applications and services for libraries, systems vendors, publishers, and users. Much of what will underpin a national strategy for monographs in the UK will be based on accurate and timely data that can help to inform and drive decision making, both locally and system wide. Further information on the NMS ideas and the strategy itself can be found on the project's webpage: www.jisc.ac.uk/research/projects/national-monograph-strategy. These examples of regional and national collection management each provide exemplars of how data is being used to help drive decision making at that system-wide level. This picture maps onto the two case studies below, which describe the innovative use of data locally to drive collections- management decisions and improve the user experience, and how local data can be aggregated to provide institutions with a national picture, helping inform local decisions in the context of a national collections landscape.

7.7 DATA-DRIVEN COLLECTIONS MANAGEMENT: FURTHER RESOURCES

Below are additional resources for individuals and institutions interested in how data and new data-driven approaches can inform collections management and development.

- **National/regional shared collections strategies and services**
  The National Monograph Strategy includes a literature/landscape review that covers a lot of relevant material around collections management and national/ international activity in this space.

- http://monographs.jiscinvolve.org/wp/2013/07/31/monographs-landscape-report/.
  More information on the Maine Shared Collections Strategy (MSCS) can be found at its website, www.maineinfonet.org/mscs. More information on the UK Research Reserve (UKRR) can be found at its website, www.ukrr.ac.uk.

- The Hathi Trust, www.hathitrust.org.
  Library game examples:

- Library Game, http://librarygame.co.uk
- ALA Game Making Interest Group Wiki, http://amemakinginterestgroup.wikispaces.com/Library+Game+Examples.

- Rice, Scott, Library Guides at Appalachian State University, http://guides.library.appstate.edu/content.php?pid=449216&sid=3680579.

7.8 USING DATA TO DEMONSTRATE LIBRARY IMPACT AND VALUE

Libraries across the different sectors, as well as archives, museums, and galleries, face unprecedented challenges, from financial and technological pressures, through to social and cultural changes more generally; and libraries, as well as academic and cultural institutions more widely, are increasingly expected to demonstrate the value they bring to students and users. Yet, the library, like other cultural heritage institutions, is often considered to be of societal or cultural value, and these are values which are difficult to measure and often resist definition by numbers.

Over the last few years libraries, and in particular academic libraries, have been developing more sophisticated and data-driven approaches to demonstrating the impact of their services and resources to the institution and beyond. During the last

two decades an increasing amount of literature has been published on the impact of libraries on their users – literature that uses data to demonstrate the value of libraries to their users. This work has explored the relationship between the library and its resources and the performance of the students. Early work took place in an environment where print was still dominant and at a time when extracting and sharing data from library and institutional systems was difficult and time consuming. Some researchers explored the wider societal impacts of libraries and demographic usage patterns.

From around 2010 onwards there was a resurgence of interest in data- driven strategy and analytics within the academic library sector. These analytic experiments share several features which differentiate them from earlier work and, to an extent, from the work taking place in other library sectors:

***Diversity of data*:** The libraries are interested in data from across the library's systems and services (gate count, e-resource usage, computer logins), as well as data from across the institution (student records, student services, registry, IT). Individually, a dataset may appear peripheral or unimportant; as part of a larger collection, each dataset becomes significantly more important.

***Actionable analysis*:** Data should not be collected for the sake of collecting it. Instead, it should contribute to a new insight or understanding that can be acted upon to improve the service or system for the user.

***Service development*:** The analysis of the data isn't just about improving the experience of users of existing services, but also about providing a basis for new types of services and interventions. These new services can be more intimately tailored to the needs of users or groups of users, thus increasing the value of the services and of the library overall.

This approach – diversity of data, actionable insights, and a focus on service development – means that these pioneer libraries are using data to drive decision making. There is recognition that effort should be invested in acting on the data for the benefit of students and users, and not in the collection of that data. Such an approach also enables new insights and discoveries to emerge; the aggregation of data can lead to new insights that the individual datasets could not yield on their own or in conjunction with one or two others. The library is becoming a critical partner in the wider, enterprise exploitation of analytics by academic and cultural heritage institutions. Here we see an emerging field of interest where the library is helping to lead the way and has a significant amount of expertise and experience to provide to the institution. The emergence of student analytics and, more embryonically, research

analytics, places the library at the heart of the analytics agenda. As analytics becomes an important strategic driver for institutions, so the library finds itself ideally placed to lead and contribute to this area. And nowhere is this expertise and knowledge more important than in the legal and ethical implications of collecting and exploiting impact data.

- ***The ethics of impact***
  One thing recognized by each of the three case studies included here is the complex and challenging legal and ethical environment surrounding the use of student and user data. We often focus on the risks of analytics, but we need to be equally clear about the risks of not using student and user data. What if a student, after failing a course, approached the institution to ask why it hadn't done everything in its power to prevent him from failing? Why hadn't the library and the institution picked up on the student's behavior and patterns? Why had there been no intervention?

If Amazon can make recommendations for books based on prior searching behavior and the purchasing behavior of others, or Google can tailor search results based on your previous searches, location, and background information, why can't the institution tailor its services to the student's particular requirements? This may take the form of relatively superficial interventions, such as recommendations for further reading, or highlighting what resources the top student in a class is reading. Equally, it may take the form of red flags for students whose behavior indicates that they have an increased risk of failure or of dropping out, or tailoring services for specific groups of students so that the services an institution provides are ultimately more equitable.

As libraries respond to the rapidly evolving information landscape, so the importance of being able to effectively gather, analyse and act on data will be ever more critical. As the three case studies below highlight, we are still very much in the infancy of this area, but the studies also demonstrate the leading role that libraries are playing in the use of data and analytics, and how this has the potential to transform the role and impact of the library in the eyes of its users and host institution. They also highlight the technical, cultural, and ethical challenges that the libraries faced in bringing this data together, and how they each overcame these barriers to change the way in which the library is viewed within the wider organization.

The three international case studies are:

- **Case Study 3.1** Stone, G., *Library impact data: investigating library use and student attainment* (University of Huddersfield), p.51

- **Case Study 3.2** Nackerud, S., Fransen, J., Peterson, K. and Mastel, K., *Retention, student success and academic engagement at Minnesota* (University of Minnesota), p. 58
- **Case Study 3.3** Cox, B. and Jantti, M., *The Library Cube: revealing the impact of library use on student performance* (University of Wollongong), p. 66.

If you'd like to find out more about the work described in this unit, and access further reading and inspiration, below are additional resources for individuals and institutions interested in demonstrating library and organizational impact and value.

More on the case studies

Library Impact Data Project (LIDP at Huddersfield), http://bit.ly/libimpact.
Discovering the impact of library use and student performance (The Library Cube), http://bit.ly/libcube.

- Library data and student success (University of Minnesota), http://bit.ly/MinnImpact.

Additional resources

- Exploiting activity data in the academic environment, www.activitydata.org.
- Library analytics bibliography, http://bit.ly/analyticsbib.
- Educause Library Analytics Toolkit, www.educause.edu/library/analytics.

7.9 SELF-ASSESSMENT QUESTIONS

Q.1 Hoh would you manage the 'national' collection in libraries?

Q. 2 Describe national/regional shared collections strategies and services with examples.

Q. 3 Define data, how would you use data to demonstrate library impact and value? Explain with examples.

Q. 4 Explain the following:

- The collections turn
- Managing the local collection
- Patron-driven acquisition
- Gamifying collections
- Library game examples

7.10 ACTIVITIES

Use existing applications and services, such as SCVNGR (www.scvngr.com), to encourage students and users to undertake tasks within the library; specifically, they are using them for library inductions.

REFERENCES

Cooper A., (2012). What is analytics? Definition and essential characteristics, CETIS Analytics Series, 1 (5). 1–10.

Kandasamy, B. P. & Benson, V. (2013). Making the most of big data: Manager's guide to business intelligence success. www.bookboon.com. http://93.174.95.29/_ads/EC133CCE54AA14A53992645E9C31BF95

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Hung Byers, A. (2011). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute.

Sedkaoui, S. (2018). Data analytics and big data. John Wiley & Sons.

Vasarhelyi, M. A., Kogan, A., & Tuttle, B. M. (2015). Big data in accounting: an overview. *Accounting Horizons*, *29*(2), 381–396.

Unit–8

GOING BEYOND THE NUMBERS: USING QUALITATIVE RESEARCH TO TRANSFORM THE LIBRARY USER'S EXPERIENCE: WEB AND SOCIAL MEDIA METRICS FOR THE CULTURAL HERITAGE SECTOR

Compiled by: Dr. Amjid Khan

Reviewed by: 1. Dr. Pervaiz Ahmad
2. Muhammad Jawwad
3. Dr. Muhammad Arif

CONTENTS

INTRODUCTION

This unit provides practical examples and techniques that have been used to gain a deeper understanding of user behaviors and motivations, when interacting online and in the physical space of institutions and libraries. At the end of the unit, self-assessment questions followed by practical activities are given to the students to gain deeper understanding of research activities and its impact on shaping both the digital and physical spaces that students and users inhibits.

OBJECTIVES

After reading this unit, you will be able to explain:

- qualitative research and the user experience
- web and social media metrics and analytics
- the social web and web impact
- user behavior: external and internal
- crowdsourcing applications and blogs
- the traditional web and web impact assessment
- social media metrics and semantic web

8.1 INTRODUCTION

This unit provides practical examples and techniques that have been used to gain a deeper understanding of user behaviors and motivations, both when interacting online and in the physical space of institutions and libraries. At the end of the unit, self-assessment questions followed by practical activities are given to the students.

The two case studies in this unit are about observing student and user behaviors, mapping the ways that they interact online and physically (with services, resources, and each other) and how they use technology and space for learning and collaborating. This section also highlights the increasing importance of this type of qualitative approach for understanding and shaping both the digital and physical spaces that students and users inhabit.

8.2 QUALITATIVE RESEARCH AND THE USER EXPERIENCE

When we talk about data helping to drive decision making in organizations, we often assume that the kinds of data we are referring to are the 'hard' numbers: the number of users, the frequency of use and so on. Yet, as organizations that deliver services and, more importantly, experiences, libraries, archives, and cultural heritage institutions are ultimately interested in understanding the behaviors, motivations and needs of users. We want to be able not only to know what users do but also to understand what their experience is like.

Much of the current interest in this type of qualitative research in cultural and academic institutions has been driven by several factors, but central to these have been the technological changes that transformed the information landscape. Innovations in user experience – driven by digital companies and organizations – have radically changed the expectations and assumptions of anyone using online (and, indeed, physical) services. Libraries and other cultural heritage institutions no longer have a monopoly on access to information neither content, nor, critically, but they are simply competing against similar or familiar organizations in providing access to that content or service. Rather, they are competing against wealthy, web-based and technology-astute companies. The rules have changed irrevocably. In a digital information environment, where the user experience is key, the distinction between services provided by libraries and the technologies of companies like Microsoft, Amazon and Google are becoming blurred or disappearing entirely. There is an urgent need for traditional information and content providers to be able to ask more nuanced and complex questions, to explore subtle variables and to seek out new and

emerging patterns in the data. To gain these kinds of deeper insight organizations need to adopt a mixed-method approach to analytics, one that 'incorporate[s] inquiries that measure and count, as well as asking open-ended, descriptive, or analytical questions' (Case Study 3.1). By the combination of both approaches the richness of the data is increased dramatically and what was once simply numbers now becomes a critical piece in a far more detailed jigsaw of user data. But we can also go further, so that what is being described when we talk about a qualitative approach is also a rebalancing of the data and analytics methodology: a rebalancing that incorporates ways that enable the user to tell their own story. The role of the librarian, archivist or curator is to listen and, when appropriate, to question the narrative; this is, after all, an active dialogue with the user, not passive listening.

These dialogues are a key ingredient in being able both to improve and refine current services and systems and to identify potential new services and meet new or unmet user requirements – often meeting users' needs that they may not have fully realized they have, or at least that they have not fully articulated.

One advantage to having a full-time ethnographer on the library's staff is that part of the value of an ethnographic approach is the possibility to be embedded in the environment over long periods of time. Short-term contract work or grant-funded research that lasts only a few years is better than nothing, but longitudinal research that is fully integrated into the everyday workings of the library allows the best chance to respond in an agile way to conditions 'on the ground'.

Read the following two international case studies are featured in this unit:
Case Study 3.1 Connaway, L. S., Hood, E. M. and Vass, C. E., *Utilizing qualitative research methods to measure library effectiveness: developing an engaging library experience* (OCLC), p. 82.

Case Study 3.2 Lanclos, D., *Ethnographic techniques and new visions for libraries* (University of North Carolina, Charlotte), p. 96

More on the case studies
Evaluating Digital Services: A Visitors and Residents Approach, www.jiscinfonet.ac.uk/infokits/evaluating-services.

Anthropology in the Stacks blog, http://atkinsanthro.blogspot.co.uk.
Additional resources

Ethnography, Usability and User Experience in Libraries (blog), http://ukanthrolib.wordpress.com.

Journal of Library User Experience, http://weaveux.org.

Ethnographic Research in Illinois Academic Libraries (ERIAL), www.erialproject.org.

8.3 WEB AND SOCIAL MEDIA METRICS AND ANALYTICS

Web metrics and analytics refers to 'the measurement, collection, analysis and reporting of web data for purposes of understanding and optimizing web usage'. As libraries, archives, galleries and museums direct greater focus and resources to the development of their online presence, so it becomes increasingly critical to capture and analyze users' online interactions and experiences. Like the institution's physical building, its web presence represents a vital part of an institution's existence. The web is now the starting point for much of what we do: finding a painting via Google search, locating an article through a link in a blog post and so on. These online interactions often lead us to the collection or item in which we are interested on the institution's website, but we may not go beyond the Wikipedia page for an object, or even the Google search results page. It's crucial for institutions to find ways to ensure that their collections and content are available in the online spaces and places that people are already inhabiting on the web; for the institution, as it were, to go where people already are.

This means that there must be a shift in the kinds of online interactions cultural institutions are willing to have with their 'visitors'. They need to learn about the changing behaviors and expectations of users and visitors, and to discover not only how they interact with the institutional web pages, but where they have come from, where they go and how they engage with the web more generally. The web has become the location for our searching, discovery, use and even creation of content; it has transformed our expectations of what content and services should be like. It is thus essential for cultural institutions to understand the changing requirements and expectations of users, and they can do this by studying what their users are doing and where they are going on the web.

Add to this the fact that the institutional web presence may no longer be a single, discrete location, but may reflect the more complex range of social spaces that users occupy across the web. The institutional web presence is now likely to be complex, dynamic and multi-faceted. Its foundation is likely to be a website, but it may be accompanied by a range of other presences, from blogs and social media

through multimedia channels. Indeed, these community and social presences are fast becoming critical components in an institution's digital existence and in its interaction, engagement and communication with its network of users and visitors.

8.4 THE SOCIAL WEB

The web is a social engine – it drives social interactions and networks. As Sir Tim Berners-Lee has written: 'The Web is more a social creation than a technical one. I designed it for a social effect – to help people work together and not as a technical toy'. The evolution of the web saw early development focused on the one-way communication of static web pages. However, further developments quickly began to enable two-way conversations and interactions, and eventually the emergence of 'social software' such as Wikipedia, where the affordances of the software are focused on the 'group', rather than the individual.

Of course, now the social web (and social software) is a ubiquitous and essential part of our web experience, from blogs and twitter to Facebook and YouTube. Institutions can no longer count on a single web presence, but instead need to inhabit multiple web spaces and, critically, engage with those spaces and the people in them in more meaningful ways. In fact, we might argue that the social web demands that institutions should behave as residents of the web, not merely visitors. With a resident approach, the institution inhabits a digital space is much the same way that you might inhabit a physical space – you see yourself as part of that community, you converse and engage and effectively live part of your life or existence in that space. Part of you continues to exist when you are not online – your persona does not disappear, as it were, when you log off. This behavior also means that we increasingly leave a trail – a 'data exhaust' – as we move across the web and interact with different spaces and networks.

The social web encourages the types of activities and interactions that produce large volumes of data – think about the number of Tweets or Likes produced in one day, globally. Getting the data is, in some ways, not the problem. Rather, the problem is what we want to measure and why.

For metrics on impact and engagement there will be the largely numerical data related to page hits, views, Likes and so forth. These are what we might describe as classic web metrics – the data telling us what people are doing, where and how often. In terms of engagement, we might also be interested in how many new people visited a website, or how many left a comment or shared a story or piece of content. These numbers are important. They enable us to compare our own

institution to other institutions in a standard way, to track our progress and to uncover new insights into our audiences and content. But impact and engagement have much deeper aspects, which institutions could and should be measuring. For example: what is the wider impact of a piece of content? How did it change or influence something – such as government policy? Why does someone like a particular blog post? – How did it improve their learning experience? Questions such as these begin to challenge and question the assumptions that underpin our current approach to web metrics. Why do we want more people to share our content? Is the size of the audience important, or do we want to get the content in front of specific audiences or people? Increasingly the web is allowing us to consider the 'bigger' metrics that can uncover the implications for us, as institutions, of our impact and engagement, and the breadth and depth of data generated through the social web challenge us to consider concepts such as engagement and impact in more nuanced ways.

8.5 WHY MEASURE WEB IMPACT?

It will quickly become clear, with so many potential web metrics available for analysis, it is essential to have a clear understanding of why the analyst is measuring web impact – if they are to pick an appropriate metric. There are eight reasons for managers within a public organization to measure performance, and they are all equally applicable to the investigation of a web presence. They are: *to evaluate, to control, to budget, to motivate, to promote, to celebrate, to learn and to improve.*

Evaluation is most of the organizations' primary motive for measuring their web impact, while an organization's desire to understand how well it is performing will not be limited to evaluating the impact of its web presence, the explicit quantitative metrics that are obtainable using new web technologies may provide more insight than qualitative indicators that rely on the organization's long-established experience or face-to-face interaction with clients.

Controlling is ensuring that employees are doing the right thing. This can include assessing, whether an appropriate amount of content is being created, or whether the right balance is being struck between the formal and less formal content that is often combined on social media.

Budgeting, whether of money or time, is an essential function in any organization, and while there are many web services and technologies that are free at the point of use (e.g., Facebook, Twitter, Tumblr), the time spent on one service is necessarily time that is not being spent elsewhere.

Motivation to achieve a particular goal can be aided by web metrics. Fuzzy concepts such as 'improve user engagement' can be replaced by specific goals that can be aimed for, such as 'add one hundred new followers.

Promotion is the use of web metrics to demonstrate the impact of a service to those outside the organization, whether they be the public, journalists or public officials. This is important for CHIs not only in the public sector, where public funding has been under increasing pressure in recent years, but also within the private sector, where information services may need to demonstrate the contribution, they are making to the wider organization.

Celebration provides another opportunity to use web metrics in measuring performance. The celebration of notable milestones, whether the 10,000th follower on Twitter or the 1,000th 'Like' on Facebook, provides an opportunity to bring people together and recognize their achievements.

Learning is about understanding why something is or is not working. Evaluation is generally the primary purpose of many web metrics, but it is not enough to know that a technology is or is not a success; it is important to understand *why* it is a success.

Improving services is our eighth reason for using web metrics. It is not enough for CHIs to understand what is or is not working, or why it is or is not working; rather, they need to understand how to change behaviors so that services are improved. However successful an organization's online presence may be, there will always be room for improvement.

In addition to these internally focused motives, web metrics may also be used to help filter the ever-increasing information deluge and to research the behavior of people who spend an increasing proportion of their lives online.

8.6 TOOL CATEGORIZATION

There are a wide range of tools and methodologies for measuring the web impact of an organizations or individual's online content. They may be broadly categorized into four types:

User behavior/user traces**:** There are tools that give insights into the way people browse or search the web, and tools that provide insights into the content they leave behind.

***Internal/external*:** There are data-collection tools that an organization incorporates into their content, and external tools that collect data automatically.

***Private/public*:** There are tools where the information is private, and tools where the information is public.

***Free/subscription*:** There are tools that are free at the point of use, and tools that require a subscription. Many of the latter have a freemium model, providing some information for free and some for a price.

8.7 USER BEHAVIOR: EXTERNAL

Comparisons of users' behavior with content from different institutions can be made by using global traffic statistics as well as behavioral metrics on social media websites.

8.8 GLOBAL TRAFFIC SERVICES

Several services provide insights into web traffic across the web: Alexa (www.alexa.com), Compete (www.compete.com), Quantcast (www.quantcast.com). These generally operate with a freemium model: some information is available for free, while the full service often requires a subscription. Each of the services not only provides access to different information but collects the information in different ways: Alexa's ranking information is based on users of the Alexa toolbar, as well as a sample of all internet users. Quantcast provides a web analytics and advertising service, collecting data from 'Quantified' sites and estimating traffic for other sites.

Unlike an internal service such as Google Analytics, external services enable comparisons between multiple institutions, although they do not generally enable the same level of detail that a website owner can get for their own site, and it can be difficult to determine the traffic for sites that don't have their own domain name.

8.9 GOOGLE TRENDS

Another source of information about the impact of institutions on the web is Google Trends. Rather than providing information about the pages people are visiting, it can provide insights into what people are searching for. This allows for insights into the impact of exhibitions or events.

8.10 SOCIAL MEDIA VIEWS

The social media services often promote a wide range of metrics, predominantly displaying information about a user's number of followers, views and the amount of content that has been created. Where an organization is sharing content, it is likely to be interested in the number of views that content has received. Facebook, YouTube and Flickr each provide view-based metrics, while social media providers rely on associated metrics such as number of followers or number of shares. Facebook provides information on the number of visitors to a page and the most visitors in a week; YouTube provides information about views per video as well as views for a whole account; Flickr provides information about how often each photo is viewed, requiring the user to aggregate this information. As with comparisons of all web content, it is important to compare like with like and to recognize the different uses to which a technology is put.

8.11 USER TRACES: INTERNAL

User traces refers to impressions left on the web deliberately, and generally these can be captured retrospectively by anyone with the necessary technology. User *behavior* must be captured at the time, either by the user (e.g., via the Alexa toolbar) or by the website (e.g., by Google Analytics), and access is controlled by the owner of the data. Whereas some data owners share some of this information widely in a suitably anonymized format (e.g. Alexa and Google), much of it is not shared. In comparison, user *traces* remain available for investigation until the trace is either changed or deleted. As with user behavior, investigations of user traces may be either internal or external; can focus either on the comments or contributions made to its own website, or on those left on the wider web. It would be a challenge to do both sufficiently well.

8.12 BLOGS

Blogs are one of the most established social media technologies, providing a simple mechanism for regularly updating a website with posts that are published in reverse chronological order, and are widely used in the cultural heritage sector. They may be associated with specific projects, specific areas of work or an institution. Importantly, blogs are not only a means for publishing information, but are a means of obtaining feedback.

Comments are the feature that distinguishes blogs from other content management systems. They enable organizations not only to show the work that

they are doing, but also to engage in conversation with users. However, even for something as seemingly simple as comments, multiple metrics may be calculated, and while each may be appropriate for a particular situation, they also have their own limitations:

- ***Number of comments***: these will vary a great deal, according to the number of posts posted.
- ***Number of comments per post*:** the raw number of comments can easily be increased by removing any moderation.
- ***Number of positive comments*:** Content analysis or automatic sentiment analysis is necessary to determine the reasons for the number of comments, and even then, the result could be adversely affected by several users posting multiple times.
- ***Number of commentators per post*:** this may be inflated by the posting of contentious issues.

8.13 CROWDSOURCING APPLICATIONS

Crowdsourcing apps are used to collect data from many users using a single app on their mobile devices. Administrators control the information and decide whether it will be accepted and visible publicly, in a crowdsourced map portal. Crowdsourcing involves obtaining work, information or opinions from a large group of people who submit their data via the Internet, social media, and smartphone apps.

Blogs are not the only way users can contribute to institutions' websites, as institutions look for ways to gain contributions from the wider community on specific projects. For example, as well as a Creative Commons project on Flickr, the British Library has also hosted its own crowdsourcing Georeference project designed to make digitized maps from the collection available in the most accessible format (www.bl.uk/maps), while the Natural History Museum is a partner in the Notes from Nature (www.notesfromnature.org) Zooniverse project to engage the public in transcribing historical records from the museum. Unlike many activities, these projects have definitive goals. Although there are many additional metrics that may be calculated (*e.g.*, number of contributors, average number of contributions per contributor) the completion of the project is the primary goal.

8.14 USER TRACES: EXTERNAL

The power of the web comes from the fact that content and websites do not exist in isolation but, rather, are interconnected. The web can be viewed as a giant conversation with people talking about and linking to the content of other

organizations. Originally these conversations happened across the web on millions of small websites, but increasingly they happen within a small number of huge social media websites with hundreds of millions of users. Both the web as a whole and large social media sites need to be considered when investigating the impact of an institution.

8.15 THE TRADITIONAL WEB

Web in general is not as consistently structured as are the profiles on social media sites, and similar information is likely to be found not only in different places on different websites, but also on different pages within the same website. This lack of consistency means that measures of impact have focused on two types of information that can be readily discerned: the text of web pages and the links between web pages.

8.16 WEB IMPACT ASSESSMENT

A web impact assessment refers to assessing the impact of ideas or documents by counting the number of times they are mentioned online. At the most basic level, the assessment may be of the number of hits a particular search gets when entered a search engine. Although web impact assessments undoubtedly have the potential to provide some insights into ideas or documents, limitations of the current web as well as the current generation of search tools means that the applicability of web impact assessments is quite limited.

8.17 INLINKS/URL CITATIONS

A level of ambiguity is inevitable with text searches, whereas the unique nature of URIs (Uniform Resource Identifiers) should leave no room for ambiguity as multiple organizations are not allowed to have the same domain name. However, the tools available for investigating the links pointing to websites (i.e., inlinks) or specific web pages are quite limited. Whereas link data was once available via the major search engines, this functionality has been depreciated by most search engines. Now such investigations must make use of the functionality provided by tools aimed primarily at search engine optimization specialists, such as ahrefs (http://ahrefs.com), Majestic SEO (www.majesticseo.com) and Open Site Explorer (www.opensiteexplorer. org). Such sites vary in terms of the amount of information that is available for free, with the finer details often restricted.

A URL citation is the appearance of a URI within the text of a page. This means that search engines will index them and can be used to investigate them. For example, "bbc.co.uk"-site:bbc.co.uk is a phrase search to find the appearance of bbc.co.uk in the text of web pages that are not part of the main BBC website, while "bbc.co.uk" site:twitter.com finds the appearance of bbc.co.uk on pages within the Twitter domain. The Webometric Analyst (http://lexiurl.wlv.ac.uk) software from the Statistical Cybermetrics Research Group at the University of Wolverhampton is designed for the collection and analysis of data from a wide range of online sources, including the search engine Bing, the last of the major search engines to allow automatic querying of its search engine. Automating the combining of URI citations on multiple web domains enables the production of network diagrams between the different websites.

8.18 SOCIAL MEDIA METRICS

Social media metrics is the use of data to gauge the impact of social media activity on a company's revenue. Marketers often use social media monitoring software to observe activity on social platforms and gather information about how a brand, product or company-related topic is being perceived. People are not just consuming content on the web, they are creating content, sharing content, and having conversations, in both their personal and professional lives. These traces provide a rich source of information for investigation and, in comparison to the diversity of traditional web pages, social network sites create a large amount of data that is structured in the same format. This has led to the creation of several desktop tools and web services available for both collecting and analyzing this data.

8.19 IMPACT OF WEB CONTENT ON SOCIAL MEDIA SITES

As well as investigating the impact of content on the web, it is also possible to investigate the impact of content on a small part of it. Both ahrefs (http://ahrefs.com) and Open Site Explorer currently provide access to social metrics about the impact of web metrics, although of the two only ahrefs (http://ahrefs.com), provides access to this data for free, and even then, only sharing the number of Google +1s, Tweets, Facebook Likes, and Facebook shares for a website's homepage. However, there are also further means of investigating impact on a specific service. In some cases, there is a simple, user-friendly web service adding additional search functionality to a site's service. For example, Topsy (http://topsy.com) provides a more extensive search service for Twitter content than is provided by Twitter itself. Many social media sites also have extensive APIs (application programming interfaces), allowing developers and researchers to

automatically download vast quantities of data. However, most people interested in investigating social media content are unlikely to have the necessary programming skills to capture this data for themselves, but there are tools that provide a more accessible user interface to the APIs. These are discussed below in the context of measuring the impact of social media content itself.

8.20 IMPACT OF SOCIAL MEDIA CONTENT

No other web content is seemingly as aligned with web metrics as that on social media sites. Users are not only encouraged to follow, share, and comment on one another's content, but the associated metrics are often prominently displayed, whether this be the number of followers on Twitter or the number of times a video has been liked on YouTube. It is important to remember, however, that not only is the nearest metric not necessarily the best, but also consideration should be given not only to the content that is captured but also the content that is not.

For instance, one of the most noticeable pieces of social content is the Facebook 'Like', enabling users to simply respond positively to content that has been put online. There is, however, no equivalent 'dislike' or 'ambivalent' button on Facebook. Users may comment negatively about the content that is shared, but this is not reflected in the simplicity of metrics that revolve around the number of 'Likes' a post has received. Equally, following someone on Twitter gives little indication of why they are followed, and retweets are not endorsements.

8.21 A SEMANTIC WEB

The Semantic Web is an extension of the World Wide Web through standards set by the World Wide Web Consortium (W3C). ... The term was coined by Tim Berners-Lee for a web of data (or data web) that can be processed by machines—that is, one in which much of the meaning is machine-readable.

As well as a seemingly endless variety of metrics for the websites and social media sites that contribute to many institutions' web presence, the ways that we use the web continue to change. In some cases, this is as Excel spreadsheets, or even PDFs, but the real power of the web comes from releasing the data in the standards of the semantic web and linking to other public datasets. This requires new tools and new metrics for understanding the impact of both the data that is published and the way that it is published.

The following two case studies present a picture of how some of the UK's biggest and most popular cultural heritage organizations (like the British Museum, British Library, Tate Gallery, V&A [Victoria and Albert Museum] and Wellcome Collection) take advantage of web metrics and analytics. The case studies provide unique insights, tips and examples of how institutions are utilizing web metrics to better understand their users' behaviors, to improve their web and digital presences and to ensure maximum impact for what they are doing on the web. The two case studies in this unit are:

Case Study 3.3 Stuart, D., *The web impact of cultural heritage institutions*, p. 117

Case Study 3.4 Malde, S. et al., *Let's Get Real: A Journey Towards Understanding and Measuring Digital Engagement*, p. 136.

8.21 SELF-ASSESSMENT QUESTIONS

Q. 1 Write a note on users' experiences with different web tools.

Q. 2 Discuss web, social media metrics and analytics with examples.

Q. 3 Define users' external and internal behaviors with relevant examples.

Q. 4 How institutions are utilizing web metrics to better understand their users' behaviors for what they are doing on the web?

Q. 5 Explain the followings.

- The social web
- Why measure web impact
- Crowdsourcing applications and blogs
- Traditional web
- Web impact assessment
- Social Media metrics
- Semantic web

8.22 ACTIVITIES

Prepare a case study project on how to take advantage of web metrics and analytics to better understand library users' behaviors, to improve their web and digital presences and to ensure maximum impact for what they are doing on the web.

REFERENCES

Cooper A., (2012). What is analytics? Definition and essential characteristics, CETIS Analytics Series, 1 (5). 1–10.

Kandasamy, B. P. & Benson, V. (2013). Making the most of big data: Manager's guide to business intelligence success. www.bookboon.com. http://93.174.95.29/_ads/EC133CCE54AA14A53992645E9C31BF95

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Hung Byers, A. (2011). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute.

Sedkaoui, S. (2018). Data analytics and big data. John Wiley & Sons.

Vasarhelyi, M. A., Kogan, A., & Tuttle, B. M. (2015). Big data in accounting: An overview. *Accounting Horizons*, *29*(2), 381–396.

Unit–9

UNDERSTANDING AND MANAGING THE RISKS OF ANALYTICS: CASE STUDY

Compiled by: Dr. Amjid Khan

Reviewed by: 1. Dr. Pervaiz Ahmad
2. Muhammad Jawwad
3. Dr. Muhammad Arif

CONTENTS

INTRODUCTION

Analytics practice is strongly linked to modern enterprise management. Users, especially born-digital generations, appear increasingly to expect personalized services that are responsive to profile, need and interest – and are therefore more likely to be content for their data to be used to those ends. In considering the collection and processing of such data, institutions need to balance risks and rewards with legal and policy obligations as well as with the expectations of their community by aligning use of personal-activity data and business intelligence with their overall mission and motives weighing the benefits and costs of putting in place policies, procedures, and tools for organizational legal and risk compliance adapting governance frameworks and developing staff awareness to cover the responsibilities related to such data taking account of capture and exploitation of student – or researcher – activity data by individual academics and service providers (both within and external to the institution) including shared services. This unit discusses the legal, risk and ethical aspects of analytics, context and contextual integrity of analytics. It also covers topics on legal, ethical issues for institutions and developing codes of conduct. At the end of the unit, self-assessment questions followed by practical activities are given to the students.

OBJECTIVES

After reading this unit, you will be able to know:

- legal, risk and ethical aspects of analytics
- context, contextual integrity, legal, ethical issues for institutions
- ethical complexity and developing codes of conduct

9.1 INTRODUCTION

There are now compelling motivations driving the development of analytics capabilities in the education sector: Responses to economic and competitive pressures may be derived from business intelligence. Analytics practice is strongly linked to modern enterprise management. Users, especially born-digital generations, appear increasingly to expect personalized services that are responsive to profile, need and interest – and are therefore more likely to be content for their data to be used to those ends.

In considering the collection and processing of such data, institutions need to balance risks and rewards with legal and policy obligations as well as with the expectations of their community by aligning use of personal-activity data and business intelligence with their overall mission and motives weighing the benefits and costs of putting in place policies, procedures, and tools for organizational legal and risk compliance adapting governance frameworks and developing staff awareness to cover the responsibilities related to such data taking account of capture and exploitation of student – or researcher – activity data by individual academics and service providers (both within and external to the institution) including shared services. Before we explore the legal and ethical risks for institutions exploiting data and analytics, it is worth briefly reflecting on the multifaceted and nuanced challenges of privacy and data protection, for libraries.

9.2 LEGAL, RISK AND ETHICAL ASPECTS OF ANALYTICS

- **Context**

As universities and colleges increasingly focus on personalized services, even the broad analytics required for library collection development and service improvement needs to take account of specific demographics (e.g., ethnic groups, modes of study) as well as more general 'borrower category' trends. Meanwhile, emergent opportunities to use library data as part of learning analytics imply a more individualized application of trend data ('People like you …'), moving away from aggregated and highly anonymized treatments. Consequently, regardless of the approaches offered in library management- dashboard systems (where aggregation typically serves the purposes of collection development), it seems prudent to review legal and ethical considerations relating to library analytics on the same basis as is necessary for the explicitly personalized intentions of learning analytics.

Personalized services, whether relating to library activity or broader learning analytics, involve the collection, storage and analysis of data on user (student, researcher, staff) behaviors and the application of such data to inform decision making and to design interventions (such as resource recommendations and early warnings), which may in turn generate more data. This data can be generated either intentionally, where the student supplies the data to meet mutually understood objectives, or unintentionally, in the form of data trails, the incidental by-product of interaction with institutional systems through website clicks, VLE accesses, library borrowing and turnstiles. However, the data is generated, the holding and use of private data on individuals is governed by legal regimes in most if not all developed societies. Here we focus explicitly on the requirements of law in respect of the key themes of data protection and consent and the related areas of freedom of information, intellectual property rights and licensing and contract law. While the letter of the law will differ in other territories, the same areas will almost certainly be relevant. In addition, the analysis of data about individuals or generated by them, and the use of such data to intervene in their activities, no matter how benevolent the intention, generates what may be regarded as ethical issues in terms of the norms that should govern the use of data, even when the use of the data is legally compliant.

Legal compliance, while necessary, is insufficient from an ethical standpoint, as it is likely to constrain the permissible use of data within a range of uses but not to dictate how it is used within that range. So, while legal compliance is a necessity for institutions using analytics, the purposes for which analytics are used and a range of issues around the terms on which they are used and the respective roles of the actors involved are an ethical matter, whether they are governed by a contract which specifies mutual obligations beyond legal regulatory requirements over the use of analytics or left to voluntary interactions. Furthermore, some commentators are concerned that certain approaches should not be treated as if they were unproblematic and unquestionable: for example, how success is defined when analytics are used to promote a goal; and the issue of whether interventions based on analytics are something that the institution and its representatives 'do' to its community of students and researchers or are to be conceived in some other way that involves active participation.

- **Contextual Integrity**

Contextual integrity is defined in terms of informational norms: it is preserved when informational norms are respected and violated when informational norms are breached. The framework of contextual integrity maintains that the indignation, protest, discomfit and resistance to technology-based information systems and practices … invariably can be traced to breaches of context-relative informational norms. Accordingly, contextual integrity is proposed as a benchmark for privacy.

- **Legal**

All data held by an institution is governed by data protection legislation, and the development of analytics raises no new issues of principle, although there may be a need to be clear as to who holds the data for the purposes of registration. New intellectual property rights may arise from the creation of new databases. The increased use and visibility of data may conceivably lead to an increase in the number of freedoms of information requests received, and the use and reuse of data, especially if it is shared between institutions, will raise licensing and possibly other contractual issues, depending on who owns the data and how they allow it to be used and by whom. Hence, in developing learning analytics programs institutions need to be aware of such issues and to ensure that policy and administrative practice are developed appropriately. However, no new issues of principle are raised and for a conscientious institution the legal risk from such work is low. The legal framework governing the collection and use of analytics requires institutional vigilance and active policy development and implementation through resourcing and processes. It implies that policy in areas such as licensing and contract management may need to be developed, but (except possibly for where the legal requirement for informed consent crosses over with ethical concerns discussed below) no new major issues of principle are raised. Because this is a new domain of activity, then, the application of the law and particular solutions have not generally been subject to legal tests in the courts, and so there is inevitably a degree of risk involved, but it would seem to be low. Given that learning analytics are normally deployed for benevolent reasons, the risk of legal challenge would also seem to be low, although this may depend on how some of the ethical issues discussed below are tackled. UK law requires that the consent of the person on whom data is being collected should always be sought; this is normally done when one first signs up to a service based on either an opt-in (you say 'yes' to data being collected) or an opt-out (data is collected unless you explicitly say 'no'). Both approaches have legal standing, although some regard opting out as legally and ethically problematic as a way of securing informed consent. Institutions will need a clear policy on this. For consent to be deemed to be proper and informed, clear, and transparent information as to what information is collected, and for what range of purposes, needs to be readily available for the user to consult.

- **Ethical**

The ethical issues surrounding analytics in education are of growing concern and the amount of literature is increasing rapidly. The following are the main ethical challenges surrounding analytics:

- Unfair discrimination.
- Reinforcing human biases.
- Lack of transparency.
- Privacy.
- Lack of transparency.
- Consent and power.

Similarly, big data analytics raises several ethical issues, especially as companies begin monetizing their data externally for purposes different from those for which the data was initially collected. The scale and ease with which analytics can be conducted today completely changes the ethical framework. We can now do things that were impossible a few years ago, and existing ethical and legal frameworks cannot prescribe what we should do. While there is still no black or white, experts agree on a few principles:

- **Private customer data and identity should remain private:**
  Privacy does not mean secrecy, as private data might need to be audited based on legal requirements, but that private data obtained from a person with their consent should not be exposed for use by other businesses or individuals with any traces to their identity.

- **Shared private information should be treated confidentially:**
  Third party companies share sensitive data — medical, financial, or locational — and need to have restrictions on whether and how that information can be shared further.

- **Customers should have a transparent view** of how our data is being used or sold, and the ability to manage the flow of their private information across massive, third-party analytical systems.

- **Big Data should not interfere with human will:** big data analytics can moderate and even determine who we are before we make up our own minds. Companies need to begin to think about the kind of predictions and inferences that should be allowed and the ones that should not.

- **Big data should not institutionalize unfair biases** like racism or sexism. Machine learning algorithms can absorb unconscious biases in a population and amplify them via training samples.

A full consideration of the ethics of analytics-based programs would need to be situated in the context of how institutions can best meet their obligations overall in the light of limited resources. Discussing the ethics of analytics in isolation can obscure the need for such all-things-considered judgements that, in practice, authorities must make. Nevertheless, the ethical issues around analytics need to be articulated, and all actors need to be made aware of the issues that such novel ventures raise.

9.3 ETHICAL ISSUES FOR INSTITUTIONS

There seems to be a consensus that institutions, and therefore their library functions, should:

- use what they know to promote student success.
- use what they know to promote their own well-being–which principally translates into promoting student retention and satisfaction.
- support students to successfully manage their own learning.
- create and maintain an environment that is conducive to academic success.
- ensure that data used in analytics is held in compliance with legal regulations, that it is as accurate and up to date as possible and that anonymization is adequate, with access to individual identifiers appropriately handled.

The main challenge to these assumptions will come from those who argue that the use of data in analytics risks breaching the right to privacy that, arguably, is at the very basis of Western liberal-democratic societies. The use of unintentionally created data robs students of the right to opt-out of data collection; and the pervasiveness of large datasets and presumptions in favor of the primacy of sociability and of sharing undermine a historically hard-won understanding that people flourish best when they have a substantial degree of privacy and solitude. While students have always been obliged to give some data to educational institutions, now the amount of data, the unwitting way in which much of it is generated and its availability for analysis and dissemination pose new challenges of control and privacy that did not arise in traditional systems.

We must learn the value of privacy and need to develop new ways of preserving it, perhaps through new commercial services. Educational institutions must therefore assess to what extent it is feasible, let alone desirable, to allow students to opt- out of at least some systems, if they wish to. At the same time, they need to do what they can to ensure that students are aware of the implications of behavior that creates data trails and of the extent to which they can maintain their privacy – consistent with the practicalities of wishing to benefit from the educational opportunities that the institution offers.

Some would want to ensure that the notion of 'student success' is not something which institutions define on behalf of students, and that it should be open to a multi-dimensional understanding: success might be getting the highest possible grade, but it might also be seen in terms of gaining understanding, or enriching cultural awareness, or building networks, or in a range of other ways. Another area of concern is the risk of bias and stereotyping. Analytics based on prior history and trends might lead to judgements (e.g., about student potential or about scope of study) which limited the ability of individuals to out-perform expectations and to develop original or serendipitous approaches to their subjects.

Furthermore, decisions based on analytics can allocate resources based on measures implicit in the analytics (e.g., achieving higher grades) rather than, for example, lead to effort being put into enriching learning for students at all levels or, more broadly, developing the library collection and its use. This can lead to an emphasis on allocating resources according to what can be measured, rather than based on a more qualitative conception of the aims of education. Some might argue that any remedial skewing of resources to favor one group of students over others goes against the presumption that each student is entitled to equal treatment, and hence equal attention, from the institution. This basic ethical claim for equity gains yet another dimension in an environment where every student is also a fee payer. However, it is easy to overlook the fact that equality is a complex concept and that it is impossible, in principle and not just in practice, to satisfy everything that it might be seen to require. As a matter of conceptual logic, equality in any one dimension inevitably means inequalities in others. Giving the disadvantaged extra help is to give unequal treatment, albeit with the aim of equalizing outcomes or opportunities. Students who pay equal fees may claim that this entitles each of them to an equal share of resources from the institution; but they may also claim that it entitles each one of them to the share of resources they need to achieve an equal outcome. And, to add further complexity, the meaning of 'equal outcome' could be read to mean either literally equal outcomes – which could imply a huge skewing of resources to help the least academically able to perform at the highest levels – or, more realistically, that each is helped to achieve according to their abilities. It is such complexity that has led many moral philosophers to conclude that the ideal which a social system needs to satisfy overall is not equality – since it must always be unequal in some dimensions even if equality is attempted in others – but *fairness*, so that the dimensions in which there is inequality are seen to be morally justified.

9.4 ETHICAL COMPLEXITY: DEVELOPING CODES OF CONDUCT

As can be seen, the ethical issues around learning and library analytics are complex and can appear different according to the perspective from which they are viewed: a business manager may well have different priorities from a faculty member or a librarian, and both may differ from students themselves. These differences need to be handled, and the full implications of analytics-based interventions, and the opportunities they offer as well as their difficulties, need to be teased out. As many commentators argue, the availability of big data in education carries with it an obligation on institutions to use it to benefit both students and the institution (within the context of an all-things-considered judgement on how scarce resources are to be allocated). Thus, the complexity of the ethical considerations involved cannot provide an excuse for inaction. Nor are these ethical considerations very different from those in which the educational enterprise is always involved. On the other hand, data needs to be used in ways that are consistent with treating students as responsible and active

agents who act in awareness of the implications of data and analytics for their educational and life chances. These considerations therefore imply a further set of ethical responsibilities for institutions wishing to take advantage of analytics to further student success, and academic excellence more broadly.

To ensure, for both moral and legal reasons, that the issues around data and analytics are presented to staff, faculty, and students as clearly and transparently as possible, institutions (and therefore libraries) should ensure that measurable dimensions of success – in particular, maximizing grades – do not swamp other considerations develop awareness of the limitations of interventions based on analytics and that they point to possibilities and probabilities rather than to certainties present analytics-based interventions in a manner that enables users to make their own choices, in particular ensuring that analytics-based interventions do not inhibit enquiry and replace learning with spoon-feeding consider how users can be involved in designing analytics-based interventions, e.g. through providing input on what data they find relevant or helpful in achieving their study goals.

As was the case for research ethics in the years following World War 2, codes of conduct for analytics need to be developed that can (a) provide a focus for making explicit many of these issues and (b) provide an opportunity to debate and accommodate the range of perspectives, thereby making productive use of diverse voices. There may be different codes for the various actors – institutions, faculty, and students. analysed policy frameworks governing learning analytics in two institutions (the UK's Open University and UNISA in South Africa) and found that they are limited to institutional concerns.

Examples of codes that cover all the actors seem still to be in the very early stages of development (e.g. https://docs.google.com/file/d/0B4jK4sS8AznvY2NoZWtt SXdPTFU/edit).

Research ethics provides a valuable basis for thinking about the issues raised by analytics and has the added advantage of recognition within the educational community. The practice adopted by leading business-to-consumer services provides a clear and legally grounded approach that is likely to be readily understood by the public in much of the world.

9.5 COMPELLING CONSIDERATIONS AND GUIDING PRINCIPLES

However, the exercise of due diligence is hampered by the speed of developments in the online world and the pressure on institutions not to be left behind in the

competition for students and for research funding. The education sector faces two issues:

- *The level of legal 'maturity'*: there is a lack of precedent to indicate the application of the law in the digital environment and therefore uncertainty remains about legal interpretation.
- *Comparable ethical settings*: bearing in mind, therefore, that practice and precedent in education are relatively underdeveloped, useful exemplars might be found in research and medical ethics and in retail and online consumer services; however, there remains an underlying question as to whether education is in some respects special.

To satisfy the expectations of the 'born digital'/ 'born social' generations, there is likely to be a need to take on ethical considerations which may run contrary to the sensibilities of previous generations, especially in respect of the trade-off between privacy and service. Notwithstanding these tensions, we conclude that there are common principles that can provide for good practice:

- ***clarity*:** open definition of purpose, scope, and boundaries, even if that is broad and, in some respects, open-ended.
- ***comfort and care*:** consideration for both the interests and the feelings of the data subject, and vigilance about exceptional cases.
- ***choice and consent*:** informed individual opportunity to opt-out or opt-in.
- ***consequence and complaint*:** recognition that there may be unforeseen consequences, and therefore provision of mechanisms for redress.

This unit provides a single case study exploring the legal and ethical risks that institutions will face in using analytics. The case study offers an overview of the current legal and ethical landscape, providing links to relevant resources and beginning to outline a code of conduct for institutions utilizing data analytics.

Case Study No. 1: Chowcat, I., Kay, D. and Korn, N., *The legal, risk and ethical aspects of analytics* (p. 157).

Additional Resources

- Cetis analytics series (several case studies on the ethics of analytics), http://publications.cetis.ac.uk/c/analytics.
- Educause Library Analytics Toolkit, www.educause.edu/library/analytics.
- Ethics, Big Data, and Analytics: a model for application, www.educause.edu/ero/article/ethics-big-data-and-analytics-model-application.
- Library Analytics and Metrics Project (LAMP) Principles, http://jisclamp.mimas.ac.uk/category/legal-and-ethical.

9.6 SELF-ASSESSMENT QUESTIONS

Q. 1 Critically evaluate the issues associated with analytics.

Q. 2 Write a comprehensive note on the legal and ethical aspects of analytics with examples.

Q. 3 What are the contextual integrity, legal, ethical issues for institutions with respect to data analytics? Discuss.

Q. 4 Explain the following:

- Developing codes of conduct for analytics
- Principles of analytics
- Context
- Ethical consideration

9.7 ACTIVITIES

As an information professional, you should draft a code of conduct for university library which cover ethical, legal, context and contextual aspects of data analytics.

REFERENCES

Cooper A., (2012). What is analytics? Definition and essential characteristics, CETIS Analytics Series, 1 (5). 1–10.

Kandasamy, B. P. & Benson, V. (2013). Making the most of big data: Manager's guide to business intelligence success. www.bookboon.com. http://93.174.95.29/_ads/EC133CCE54AA14A53992645E9C31BF95

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Hung Byers, A. (2011). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute.

Sedkaoui, S. (2018). Data analytics and big data. John Wiley & Sons.

Vasarhelyi, M. A., Kogan, A., & Tuttle, B. M. (2015). Big data in accounting: an overview. *Accounting Horizons*, *29*(2), 381–396.

_____[ ]_____